DYNAMIC ROBUST SPARSE MODELING AND SAMPLING OF HIGH-DIMENSIONAL DATA STREAMS FOR ONLINE MONITORING AND CHANGE DETECTION

A Dissertation
Presented to
The Academic Faculty

by

Mohammad Nabhan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August, 2019

DYNAMIC ROBUST SPARSE MODELLING AND SAMPLING OF HIGH-DIMENSIONAL DATA STREAMS FOR ONLINE MONITORING AND CHANGE DETECTION

Approved by:

Dr. Jianjun Shi, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Nagi Gebraeel
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yajun Mei, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kaibo Liu
Department of Industrial and Systems Engineering
University of Wisconsin-Madison

Dr. Kamran Paynabar
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: April 17, 2019

ACKNOWLEDGEMENTS

I wish to express my heartfelt thanks to my advisors and mentors, Dr. Jianjun Shi and Dr. Yajun Mei for their continuous support and invaluable wisdom. They have fostered my ability to identify and pursue substantial research topics with practical implementation to contemporary engineering issues. It has been a pleasure and an honor to work under their supervision and insightful guidance. Without their patience, dedication, and unfaltering faith in me, the completion of this thesis would not have been conceivable. A special thanks to Mrs. Shi for her generosity and kindness.

My gratitude also goes out to my thesis committee members, including Dr. Kamran Paynabar, Dr. Nagi Gebraeel and Dr. Kaibo Liu. Their constructive criticism, enriching discussions, and valuable instruction have been invaluable assets during my PhD. study. I would like to also thank them for accommodating my request for them to join the committee.

Thank you to all my professors in King Fahd University of Petroleum and Minerals for preparing me for this great undertaking during my undergraduate studies. In particular, I wish to give due recognition to Dr. Shokri Selim, who has been incredibly supportive before and throughout my PhD. study. I would like to extend my thanks to Professor Fouad Al-Sunni, who encouraged me to pursue higher education upon the completion of my undergraduate studies. My sincerest thanks to Dr. Salih Duffuaa and Dr. Mazen Elrougi, for their support and recommendations during my search for PhD. programs.

Finally, and most importantly, I wish to thank my mother, sister, uncle and aunt for their unconditional love and unending support. They are the fuel for my motivation and without their help I would not be where I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The advent of advanced sensing technology in manufacturing and service systems created data rich environments with complex structures. This resulted in unprecedented opportunities for online system monitoring and change detection. However, the challenges imposed by the properties of such environments requires novel approaches that address them in order to achieve effective process control. These challenges include the high-dimensionality of the data streams, limitations on available resources, complex imbedded structures, outlier observations, and time varying system settings, among others.

This thesis contributes to the area of System Informatics and Control (SIAC) to develop systematic and dynamic methodologies for effective monitoring and change detection in complex systems. The proposed procedures facilitate (1) dynamic strategies for sampling in the event of resource constraints, (2) robust modelling complex data structures with sparse spatial dependencies, (3) adaptive updating of system models based on novel features extracted from online observations. This thesis ties advanced statistical methodologies and engineering knowledge to address practical applications in various areas such as advanced manufacturing service systems.

The research begins with addressing manufacturing and service systems with resource limitations. In Chapter 2, we investigate methods for sampling within data rich environments and propose a dynamic sampling strategy for monitoring these environments with restricted resource. A procedure called “Correlation based Dynamic Sampling” (CDS) that leverages spatial dependencies within the data streams to improve decision making when deploying sensors in real time. Chapter 3 examines the system modelling

aspect of data rich environments by exploring dimension reduction methods. We develop a dimension reduction method named "Robust Sparse Principal Component Analysis" (RS-PCA), that is designed to robustly estimate a lower dimensional subspace by exploiting the sparse structure that is typical in high-dimensional data. The probabilistic approach for modelling offers a direct medium for making inferences on system conditions. Subsequently, Chapter 4 extends the aforementioned RS-PCA procedure for implementation in dynamic systems. The proposed adaptive RS-PCA method reduces the false alarm rate that may result from implementing static procedures. The trade-off between learning from novel observations and overfitting is managed by proposing an adaptive robust learning rate through stochastic variational inference.

In summary, this thesis sheds the light on the challenges of modeling and sampling within data rich environments for the purpose of process control. Adaptive systematic modelling and sampling strategies are developed to address common challenges from these environments. Furthermore, these strategies are implemented on several exemplary systems to assess their capability for real time application in practical scenarios.

CHAPTER 1. INTRODUCTION

This chapter will serve as an outline for the remainder of this thesis. It begins by motivating the research before identifying the objectives. A brief overview of the current state-of-the-art followed by an illustration for the organization of the thesis.

1.1 Motivation

Current manufacturing and service systems are capable of generating an overwhelmingly large amount of data in real time throughout all stages of operation. This is largely due to the incredible development in sensing technology as well as the ubiquitous and abundance use of sensors. While the resulting plethora of information likely hold key performance measurements and indicators of process efficiency and capability, identifying and extracting these informative features is extremely challenging at best with current benchmark techniques. These challenges originate from the complex data structure that is prevalent in modern high dimensional processes. Key issues include, but are not limited to, data transmission and processing capacity, sparse correlations, presence of outliers and time varying operating conditions. To address these issues, the development of methodologies that facilitate (1) the dynamic assignment of available sensing resources to circumvent limitations with minimal loss of information, (2) the exploitation of embedded sparse spatial structure to robustly reduce the overall dimension while maintaining interpretability, (3) the adaptive updating for time varying processes. On the basis of these initiatives, this thesis aims to develop systematic procedures for effective system modeling, processing monitoring, change detection and fault diagnosis for overall enhanced system performance.

1.2 Research Objectives

This research focusses on addressing crucial challenges that commonly arise when dealing with manufacturing and service systems that generate data rich environments. To this end, this thesis proposes:

i. Establishing a dynamic sensor deployment strategy for monitoring systems with limitations to transmission or computing capacity.
ii. Developing a dimension reduction technique that robustly estimates the sparse subspace with intuitive application to monitoring and diagnosis;
iii. Developing an adaptive modelling that can make online adjustments for time varying processes;

1.3 State-of-the-art

The advent of data rich environments is the outcome of the ubiquitous use of multiple sensors in modern manufacturing and service systems. On one hand, this advancement made way to unprecedented opportunities for process monitoring. On the other hand, the accompanying challenges of handling this type of data with its inherent complexity have yet to be fully explored. Modern quality control techniques may suffer when applied directly to these environments, especially due to the need of real time process monitoring and change detection. These data rich environments include, but are not limited to imagery and video capture devices, Distributed Sensing Networks (DSNs), and big data.

The most straightforward way of dealing with high dimensional data is the reduction of the dimension and transforming the original subspace into a much smaller subspace.

This allows for the use of existing techniques on the new subspace that would have been ineffective or even impractical on the original subspace. The most popular dimension reduction technique that is currently used as a common reference point is principal component analysis. However, PCA been shown to produce extremely inconsistent estimates in high dimensional settings, when the low dimensional space is sparse (Ma 2013). Not to mention the inherent issues that arise from poor interpretability as the estimated principal components are linear combinations of all the data streams (Archambeau and Bach 2009, Guan and Dy 2009).

Furthermore, limitations arise when attempting to monitor such complex systems that produce dense data in real time. Capacity constraints in the data acquisition, transmission and integration are few examples of these limitations. In practical scenarios, the only solution is to view or process a fraction of the full observations, and make decisions based on this partial information. While DSN studies the sensor layout, the allocation is fixed. Therefore, shifts that occur outside the chosen sensors will not be readily detectable. For a more in-depth review of developments in DSNs, refer to Studies in (Mandroli et al. 2006, Ding et al. 2006). To improve the monitoring effectiveness, several adaptive sensor assignment methods have been introduced. Most recently (Liu et al. 2015a) proposed an adaptive sampling strategy that is based on the Top-r CUSUM procedure first introduced in (Mei 2010). Although their proposed Top-r adaptive sampling strategy (TRAS) addresses the dynamic allocation of resources, the interdependencies of the data streams is neglected in the analysis, which can cause issues when applied to systems with strong embedded correlation structures.

1.4 Organization of the Thesis

This thesis is organized in a multiple manuscript format. Each of Chapters 2, 3, and 4 are written as a research paper, which was either submitted or is in preparation for submission for journal publication.

In Chapter 2, a systematic dynamic sensor selection strategy is proposed with the aim of online monitoring of high-dimensional data streams in the event of limited resources. The developed monitoring scheme exploits the embedded spatial structure for the purpose of making more educated inferences on the overall state of the system based on the partially observed spectrum of the original data streams. The procedure is evaluated by applying it on a real solar flare monitoring data set as well as a study of in-line Raman spectroscopy for carbon nanotube manufacturing. The integration of the partial information with the estimated (or know from engineering knowledge) interdependencies of the data streams has several advantages over other approaches that assume independence and solely uses the partial information. These advantages include better dynamic allocation of resources in subsequent acquisition times, reduced detection delay and faster conversion to expected fault location.

In Chapter 3, we start our investigation by examining the dimension reduction route for addressing data rich environments. With the aim of eventually using the reduced subspace for process monitoring, change detection and diagnosis, we adopt a probabilistic approach for our dimension reduction procedure. The utilization of a probabilistic model creates an intuitive transition into inferences over the estimated parameters. These inferences can then be used for making decisions on the state of the system. We develop a

probabilistic PCA method that is designed to robustly estimate the subspace by exploiting the sparse structure that is common in high-dimensional data. The sparse and robust variations of the standard PCA allow for more accurate and consistent estimation, which is further explored with simulation studies. To validate its efficacy in a practical scenario, we test the performance of our "Robust Sparse Principal Component Analysis" (RS-PCA) in change detection for in-line Raman spectroscopy.

In Chapter 4, we extend the work of Chapter 3 to time varying processes that require adaptive online modelling techniques. These adaptive methods can reduce the false alarm rate that would otherwise result from a static representation of the operating conditions. One challenging aspect of having an adaptive modelling technique is to handle the trade-off between adjusting the model based on novel observations while avoiding overfitting issues, especially in the event of outliers. We propose an adaptive implementation of the procedure discussed in Chapter 3. This is achieved by using stochastic variational inference for solving the probabilistic robust sparse principal component analysis (RSPCA) model of original static model. We test the performance of the adaptive method by using it to detect faults for in-line Raman spectroscopy under time varying operating conditions.

Finally, Chapter 5 provides a summary of the topics discussed in this thesis. Additionally, it provides some final thoughts as well as a look into future topics that may be explored on the basis of the discussions of this thesis.

CHAPTER 2. CORRELATION BASED DYNAMIC SAMPLING FOR ONLINE HIGH DIMENSIONAL PROCESS MONITORING

2.1 Introduction

The ubiquitous use of sensing systems in manufacturing, healthcare, biosurveillance, network security, and service processes has created data rich environments that have presented challenges for real-time monitoring and analysis. This is especially true in the environments with limited resources, whether at the data acquisition level or processing level. For instance, when low-cost wireless sensor networks are employed to monitor volcanos (Pereira et al. 2014), one might want to prolong the lifetime of such network by turning on only a limited number of battery-power sensors unless the volcano is active. When using a touch-probe coordinate measuring machines (CMM) to monitor wafer manufacturing processes (Jin et al. 2012), the current profile measurement schemes are time-consuming. Therefore, it is essential to reduce the number of samples measured in wafers while still providing an adequate process quality monitoring. Besides physical devices, the term "sensor" can also be used to denote any sources that generate relevant information. Moreover, in many real-world data rich environments, we often also face resource constraints in the capacity of acquisition, transmission, analysis, or fusion of data. In biosurveillance and epidemiology, when the Center for Disease Control and Preventions (CDC) monitors the drug resistance of certain infectious diseases such as gonorrhea, it has limited capacity for drug resistance tests. Thus, it is crucial to decide how to effectively allocate the resources to monitor which kinds of affected patients, sub-populations, or regions. Hence, in the general context of real-time or online monitoring high-dimensional

data streams in resource constrained environments, it is important to dynamically sample those informative local data streams while making adequate online anomaly detection.

There are several recent articles that tackle this problem by introducing an adaptive sampling scheme that is capable of making inferences on the state of a system in real time using a fraction of the full observation spectrum. Liu et al. (2015a) proposed an adaptive sampling strategy with resource limitations, where data streams are assumed to be normally distributed. Furthermore, a nonparametric adaptive sampling procedure under limited resources has been proposed by Xian et al. (2018b). These methods assume that the data streams are spatially independent, which means that observations collected from different sensors at any given time are independent. Wang et al. (2017) proposed an adaptive sampling strategy that considers spatiality by assuming that faults occur in a local area. However, the data streams are still assumed to be spatially independent both before and after the occurrence of a fault. The development of these existing methodologies relies heavily on the spatial independence assumption by monitoring local streams individually, and it is unclear how to extend them to more complicated data models.

In this chapter, we apply the ideas of the celebrated Upper Confidence Bound (UCB) algorithm proposed by (Lai (1987), Lai and Robbins (1985)) in the Multi-Armed Bandit (MAB) problems to Statistical Process Control (SPC), and develop effective process monitoring of high-dimensional data streams with embedded spatial structure for environments with limited resources. In many real-world applications of SPC, the anomalies are often clustered and sparse, and thus we need to balance the tradeoff between randomly searching for possible anomalous local data streams or local regions (exploration) and performing focused sampling on local data streams or local regions near

the anomalous regions for quick detection (exploitation). Now the exploration-exploitation tradeoff has well-studied in the MAB problems, and the key idea of the celebrated UCB algorithm is to use the upper confidence bound of the parameter estimation for adaptive sampling. These inspire us to explore the embedded spatial structures of local data streams/sensors to use the upper confidence bound of the local stream post-change parameter estimator to develop efficient dynamic sampling methods for online monitoring and SPC. It turns out that the existing method in Liu et al. (2015a) is a special case of our proposed methods for independent data, and thus is a UCB-type algorithm for SPC. We feel that our combination of MAB and SPC is novel, and this opens a new research direction in SPC for dynamic sampling of incomplete high-dimensional data monitoring under resource constrained environments.

We should acknowledge that dynamic sampling strategies in SPC literature usually revolve around the temporal domain where the objective is mainly to inspect the quality of the product or service (Montgomery 2009). In such scenarios, the limitation is in the frequency of acquisition times, which is usually associated with the cost of data acquisition. A common example of the cost of acquisition is when the quality inspection procedure calls for a destructive test on the parts being produced. Meanwhile, our sampling strategies are over the spatial domain, and the issue lies in the capacity of deploying, observing, transmitting, or fusing all the available sensors that are monitoring the process at any given time. The key concern that we address is how to utilize the information embedded in the spatial structure of the data streams to improve the effectiveness of the monitoring procedure. This allows for a more informative and intuitive framework when dynamically sampling the partition of streams to be observed at any given acquisition period.

A dynamic sampling strategy based on the correlation structure of data streams is characterized by how it accomplishes the following tasks at every data acquisition time t: (1) determining the fraction of sensors to be deployed; (2) providing an educated compensation for unobserved readings of undeployed sensors based on their correlation with measured variables from deployed sensors; (3) computing local statistics for deployed sensors based on the observed measurements while using the correlation based compensations for the undeployed ones; (4) fusing these local statistics into a single global statistic for global-level decision making.

The novelty of our proposed dynamic sampling method lies in exploiting the spatial correlation structure to provide an upper confidence bound of post-change parameter estimation, and is therefore named Correlation based Dynamic Sampling (CDS). The procedure is dynamic in both the sampling process of the variables to be observed at each acquisition period, as well as in providing compensation for the unobserved partition. The dynamic behavior is achieved by combining the correlation structure with the information obtained from the observed partition of the data streams. The dynamic compensation we propose is constructed from the upper confidence bound of the marginal conditional distribution of the unobserved variables given the observed variables. When a well-structured framework such as multivariate normal distribution is assumed, the marginal conditional distribution is very well defined to be another Gaussian distribution. The marginal distribution is tractable even in high-dimension when the spatial structure is readily available. This sensor assignment procedure allows for a pseudo-random sampling strategy when the process is in-control, as well as fast localization of faulty variables when the process is out-of-control. We use the term “pseudo-random” here because although the

sampling procedure tends to select a cluster of variables to be observed at any given time based on the spatial structure, the clusters themselves are randomly constructed. Furthermore, these clusters are formed from variables that are correlated, this feature of cluster formation will be illustrated further in the simulation and case studies.

The remainder of this chapter is organized as follows: In Section 2.2, we provide a brief review of topics in the literature relevant to the issue of limited resources, followed by a more detailed overview of adaptive sampling methods in the literature. Next, in Section 2.3, we discuss in detail our proposed adaptive sampling strategy for online high-dimensional process monitoring. It also illustrates two properties pertaining to the behavior of the sampling procedure depending on the state of the system. Section 2.4 assesses the performance of our proposed sampling strategy on virtually simulated scenarios, while Section 2.5 tests the performance using two case studies: one is solar flare detection (2.5.1) and the other is in-line Raman spectroscopy (2.5.2). We then finally conclude the chapter with a brief discussion of the key findings of our proposed monitoring scheme.

2.2 Relevant Topics and Review on Adaptive Sampling

The following section is split into two further sub-sections. The first (Sub-section 2.2.1) provides a brief review of relevant topics that address different aspects of resource limitations from our problem, whereas the second (Sub-section 2.2.2) gives an overview of closely related procedures discussed in the literature as well as the renowned UCB algorithm in the classical multi-armed bandit problem. This will lay the proper foundation for our discussion later.

2.2.1 Topics on Limited Resources

There are two main problems explored in the literature that share some resemblance to our limited resources process monitoring setting from an application perspective. The first being optimal design of sensors in a DSN system. In a DSN, the objective is to find a fixed sensor layout optimized for process monitoring. However, due to the fixed layout, shifts that occur outside the predefined layout will reduce the detection power, as well as the diagnostic capability, as discussed in (Li and Jin 2010, Liu and Shi 2013). Studies in (Mandroli et al. 2006, Ding et al. 2006) provide inclusive reviews of the state-of-the-art advances in DSNs for enhancement in quality and productivity.

The theory of searching and tracking targets is another application that bares resemblance to our research problem. The objective of studies in this application area is to obtain an effective employment of the limited resources available to locate a target object of interest that is within an unknown location (Frost and Stone 2001, Lim et al. 2006, Zoghi and Kahaei 2010, Ben-Gal and Kagan 2013). The main assumption of these studies is that there exists a singular object in the searching space but at some unknown locations.

This chapter differs from the aforementioned applications in that our objective is to develop a *dynamic* monitoring strategy where the data streams are correlated and are flowing continuously with the uncertainty that a failure, target, or event may or may not occur to the system. Furthermore, in its core, our proposed methodology does not assume prior information on the failure characteristics. Nonetheless, it is also capable of incorporating such information seamlessly as will be demonstrated in the case studies in Section 5.

2.2.2 Adaptive Sampling Methodologies

There are several ways of approaching the issue of monitoring a process with limited resources. The two most forward approaches are (i) random sampling and (ii) choosing a fixed set of variables to monitor. While both of these approaches can be effective in certain situations, they both suffer from not utilizing any information gained during the monitoring procedure. For example, setting fixed sensors can only detect changes in the sensors selected, but it is rare in practice to have perfect knowledge about where the fault may occur. On the other hand, while random sampling might eventually detect a change in a subset of sensors, its detection delay can be large if the magnitude of the change is not large enough to set an immediate alarm, as the process switches to monitor a different set of sensors in the next acquisition period.

One of the most relevant and recent research efforts has been done by Liu et al. (2015a) who proposed an adaptive sampling strategy that is effective for the online mentoring of high-dimensional data streams. Their proposed method was based on a procedure called Top-r CUSUM, which was first introduced in (Mei 2010). Although their proposed Top-r adaptive sampling strategy (TRAS) was shown to be effective for online monitoring of high dimensional data streams, it is limited to applications where there is no significant embedded correlation structure in the streams and independence across different data streams can be assumed. Furthermore, a similar adaptive sampling procedure under limited resources has been proposed by Xian et al. (2018b). Their method is a

nonparametric approach that addresses a similar problem under the independent assumption except when the underlying distribution of data streams is unknown.

The aforementioned proposed algorithms in the literature monitor individual sensors or local data streams by computing local statistics based on the commonly used cumulative sum (CUSUM) procedures in statistical process control, and then take advantage of the independence assumptions across different sensors to construct the global monitoring statistic based on the sum of a few larger local CUSUM statistics. These methods address the limitation of resources by assigning a uniform non-informative constant compensation value to all the undeployed sensors.

Wang et al. (2017) proposed an adaptive sampling strategy under the assumption that data streams are spatially independent, and the occurring faults affect a local cluster of sensors within a grid. The method requires setting the cluster size, which typically would require former knowledge of fault patterns. Another study on climate simulation (Xian et al. 2018a) attempts to address the challenge of dynamically sampling data and deciding which to archive due to memory limitations. However, this problem is different than ours in the sense that the limitation is not in acquiring the data, but rather in choosing what is worth keeping.

Next, we provide a brief review on the classical multi-armed bandit problem (MAB), which includes many useful adaptive/dynamic sampling methodologies. In the classical MAB, one assumes that there are p sensors or arms, and the i-th sensor generates observations over time, say, $\{X_{i,1}, \cdots, X_{i,t}, \cdots\}$, which are i.i.d. with (unknown) mean μ_k. At each time step t, one can take observations only from one sensor, say, the $i^*(t)$-th

sensor. This is equivalent to the case when one takes the observation $X_{i^*(t),t}$, and under the simplest formulation, one receives the reward $r_t = X_{i^*(t),t}$. Then one wants to decide which sensor to take observation at each and every time step so as to maximize the expected overall rewards, $E(\sum_{t=1}^{T} r_t)$. An intuitive appealing and myopic policy is to estimate each unknown mean μ_k by the corresponding sample mean of each sensor, and then take observations from the sensor that has the largest sample mean. Unfortunately such a myopic policy performs poorly and is sub-optimal. Meanwhile, one of the asymptotically optimal policies is the notable UAB algorithm proposed by (Lai (1987), Lai and Robbins (1985)). The main idea of the UAB algorithm is to choose the arm with the largest upper confidence bound estimator of μ_k. This is achieved by taking into account both the current point estimate of μ_k and the number of samples taken from each sensor, which will balance the tradeoff between exploration-exploitation. Here we extend the classical MAB with two twists: one is the changing environments with a different reward function on quick detection, and the other is to increase the number of sampled sensors from 1 to $m \geq 1$. We propose to apply the idea of the UAB to SPC, which leads to a dynamic compensation value to the unobserved streams or sensors based on the spatial correlation structure of the data and the information obtained from the observed streams or sensors.

2.3 Correlation-Based Dynamic Sampling (CDS) Strategy

In this section we develop a method for effective monitoring of correlated high-dimensional data streams under the constraint of resource limitations. In our proposed strategy, we first construct efficient local statistics for each individual data stream and consequently combine these local statistics into a single global statistic while utilizing the

information embedded in the correlation structure of the streams. There are two novel ideas in the proposed strategy: (1) following the multi-armed bandit algorithm to explore the spatial correlation structure and introduce a dynamic compensation value for the unobserved variables based on the confidence limit of their parameter estimates, and (2) deploying sensors to those variables smartly so as to collect as much global change information after adjusting the spatial correlation.

The following subsections will elaborate on the steps of our proposed correlation based dynamic sampling (CDS) strategy. Section 2.3.1 provides a detailed overview of our algorithm. Next, a detailed discussion of parameter settings is provided in Section 2.3.2. Finally, Section 2.3.3 discusses options for estimating or imposing the embedded spatial structure of the data streams.

2.3.1 CDS Methodology Development

In preparation to our discussion, we will first introduce the notations for the variables that will be used throughout the course of this chapter. Suppose that the system to be monitored consists of p variables $\mathcal{P} = \{1, \dots, p\}$ that can be observed at any time t. The vector of observed variables at time t is given by $\boldsymbol{X}_t = (X_{1,t}, \dots, X_{p,t})'$. Due to limitations in the resources available for monitoring, only a fraction of this vector is actually measured in real time. Let m be the maximum number of variables/sensors that can be measured/deployed at any acquisition time. From the problem statement, m is a process parameter dictated by the system monitoring capability. This could translate to the number of sensors available for deployment at each acquisition time, the transmission capacity, or the computational power at the data fusion center. To facilitate referencing

measured variables at each time t, we introduce two sets $\omega_t \subset \Omega$ and $\psi_t \subset \Psi$. Here, Ω and Ψ are all possible partitions of the data streams into observed and unobserved sets, respectively. Thereby, variable $X_{i,t} \in \omega_t$ if and only if it is measured at time t, otherwise it is assigned to set ψ_t. Hence, the cardinalities of ω_t and ψ_t are respectively $|\omega_t| = m$ and $|\psi_t| = p - m$.

We assume that $\boldsymbol{X}_t$ comes from a multivariate normal distribution, where the mean vector is μ_t and the covariance structure Σ. The covariance structure plays an important role in our proposed dynamic sampling procedure. Particularly, the covariance between the unobserved sensors and the observed ones (denoted by $\Sigma_{\Psi\Omega}$) is the base of inferences to be made on unobserved sensors. The in-control mean and covariance are also assumed to be known. In reality, while these parameters are not generally known, they can be estimated from an adequate amount of historical data. They can also be set to target values defined by the engineering design of the process. Without loss of generality, we assume that the data has been preprocessed to have mean 0 and standardized to have a covariance matrix equal to that of the correlation matrix. After some point in time τ during the operation of the monitored system, a change in the mean vector occurs, where a subset Θ of the variables $\boldsymbol{X}_t$ will have a non-zero mean. Moreover, we assume that the correlation structure remains unchanged during this change and thereby stationary throughout the whole monitoring period. Our objective then becomes to first detect this change with minimum delay from the onset at τ. Secondly, we need to identify the subset Θ with the shifted mean, while only observing a fraction of the variables at each acquisition time.

There are four components to our proposed method. First, we construct the local statistics for the deployed sensors based on the observed measurements. Second, we utilize the correlation between undeployed sensors and deployed ones to determine the local statistics of the unobserved variables. Third, we select the fraction of sensors to be deployed at the next acquisition time. Finally, we fuse the local statistics into a multivariate global statistic that is used to test whether or not the process remains in-control. In the following subsections we will demonstrate how we can construct each one of the aforementioned components and then conclude with an overview of the proposed monitoring scheme.

2.3.1.1 Determining Local Statistics

Our objective in this chapter is to detect any change to the mean of the monitored variables. Since this shift can be either positive or negative, it is appropriate to deploy a two-sided CUSUM monitoring statistic for each variable k at time t defined as

$$C_{k,t} = max(C_{k,t}^{+}, C_{k,t}^{-}), \tag{2.1}$$

where the notations $C_{k,t}^{+}$ and $C_{k,t}^{-}$ represent, respectively, the positive and negative local statistics for variable k at time t.

At any given time, we are limited by the available resources, computation power, or transmission capabilities to calculate the aforementioned local statistics using partial observations. Statistic pertaining to an observed sensor $X_{k,t} \in \omega_t$ at time t can be defined as CUSUM statistics (Lorden (1971)) as follows:

$$C_{k,t}^{+} = max\left(0, C_{k,t-1}^{+} + \delta X_{k,t} - \frac{\delta^2}{2}\right) \text{ and } \quad C_{k,t}^{-} = max\left(0, C_{k,t-1}^{-} - \delta X_{k,t} - \frac{\delta^2}{2}\right), \tag{2.2}$$

where $C_{k,0}^{+} = C_{k,0}^{-} = 0$. Here, δ is the smallest mean shift magnitude that is of interest to detect (see the guidelines in sub-section 2.3.2 on how to determine the value of δ).

The main difficulty is how to define the local CUSUM statistics in (2) for those unobserved variables $\left(X_{k,t} \in \psi_t\right)$. Inspired by the UAB algorithm of (Lai (1987), Lai and Robbins (1985)) for MAB, here we propose to salvage (2) by utilizing the spatial correlation structure to obtain the estimated upper and lower bounds, say, $U_{k,t}$ and $L_{k,t}$, on the true unobserved variable $X_{k,t}$ at time t (the estimates of $U_{k,t}$ and $L_{k,t}$ will be discussed in a little bit). Then we dynamically construct the local statistic as follows:

$$C_{k,t}^{+} = \max\left(0, C_{k,t-1}^{+} + \delta U_{k,t} - \frac{\delta^2}{2}\right), \tag{2.3}$$

$$C_{k,t}^{-} = \max\left(0, C_{k,t-1}^{-} - \delta\, L_{k,t} - \frac{\delta^2}{2}\right). \tag{2.4}$$

It remains to discuss how to obtain the estimates, $U_{k,t}$ and $L_{k,t}$, for unobserved sensors $\left(X_{k,t} \in \psi_t\right)$. Since the data streams are assumed to come from a standardized multivariate normal distribution, the marginal conditional distribution of an unobserved variable $X_k \subset \Psi$ over the remaining set of observed variables Ω is also normal with mean μ'_k and variance σ'_k given by:

$$\mu'_k = \Sigma_{k,\Omega}\Sigma_{\Omega\Omega}^{-1}X_{\Omega}, \tag{2.5}$$

$$\sigma'_k = 1 - \Gamma_{kk}, \tag{2.6}$$

where, $\Sigma_{k,\Omega}$ is the covariance between $X_k \in \Psi$ and the observed variables in Ω. Moreover, Γ_{kk} denotes the k^{th} diagonal entry of $\Gamma = \Sigma_{\Psi\Omega}\Sigma_{\Omega\Omega}^{-1}\Sigma_{\Omega\Psi}$.

Using the marginal conditional distribution of an unobserved variable $X_k \in \Psi$, we can construct an $(1-\alpha)100\%$ two-sided confidence interval as follows:

$$CI_{k,t} = \left[L_{k,t}, U_{k,t}\right]$$
$$\text{where,} \quad L_{k,t} = \mu'_{k,t} - \Phi^{-1}(1-\alpha/2)\sigma'_k \;\; and \; U_{k,t} = \mu'_{k,t} + \Phi^{-1}(1-\alpha/2)\sigma'_k. \tag{2.7}$$

Here, $\Phi^{-1}(.)$ is the inverse of the cumulative standard normal distribution. Hence, the bounds of the confidence interval $U_{k,t}$ and $L_{k,t}$ will be the base of our correlation based dynamic compensation procedure given in equations (2.3) and (2.4).

It is informative to compare our proposed dynamic compensations in (2.3) and (2.4) with the static uninformative compensation in (Liu et al. 2015a). In that study, the local statistics for an unobserved variable are based on a static compensation $\Delta \geq 0$, and are defined as follows:

$$C^+_{k,t} = C^+_{k,t-1} + \Delta \text{ and } C^-_{k,t} = C^-_{k,t-1} + \Delta. \tag{2.8}$$

However, Liu et al. (2015) did not provide any statistical justification why one needs to add a static compensation Δ for unobserved variable.

The following proposition shows that the method in (Liu et al. 2015a) is a special case of our approach for independent data streams, and thus the compensation defined in equation (2.8) is essentially an upper bound confidence (UCB)-type algorithm in the SPC context.

Proposition: Our proposed dynamic compensation procedure is a generalization of the constant compensation, and is consequently equivalent to it when all data streams are spatially independent. In that case, $\Delta = \delta\Phi^{-1}(1-\alpha/2) - \frac{\delta^2}{2}$.

Proof: For spatially independent data, and for any partition of the data into observed and unobserved sets Ω and Ψ, the covariance between the two sets $\Sigma_{\Psi\Omega} = 0$. Consequently:

$$\mu'_k = 0,\ \sigma'_k = 1,\ CI_k^{\pm} = \pm\Phi^{-1}(1-\alpha/2) \text{ for all } \{k: X_k \in \psi\},$$

$$C_{k,t}^{+} = \max\left(0, C_{k,t-1}^{+} + \delta\ \Phi^{-1}(1-\alpha/2) - \frac{\delta^2}{2}\right),$$

$$C_{k,t}^{-} = \max\left(0, C_{k,t-1}^{-} + \delta\ \Phi^{-1}(1-\alpha/2) - \frac{\delta^2}{2}\right).$$

Let $\Delta = \delta\Phi^{-1}(1-\alpha/2) - \frac{\delta^2}{2}$ and choose (δ, α) such that $\Delta \geq 0$. Then, the update reduces to the format in equation (2.8). ■

The main reason that the confidence limits, $U_{k,t}$ and $L_{k,t}$, are chosen to represent unobserved instances rather than the middle of the confidence interval is to promote exploration during the in-control phase of the process by favoring those unobserved variables that have been sampled less. It can be noted that a compensation based on the middle of the interval would require the significance level $\alpha = 1$, and yields a myopic policy that only uses the estimated means for decision making. Moreover, when an unobserved variable is independent of all observed variables, the previous proposition suggests that the compensation $\Delta = \frac{\delta}{2} - \frac{\delta^2}{2}$, which might be negative. A negative compensation will result in a monotonic decrease in the local CUSUM statistics until they hit 0, which will in turn diminish the likelihood of those variables to ever be explored at future acquisition times. Further discussions of appropriate parameter settings and their role in promoting the in-control variable exploration behavior are available in subsections 2.3.1.5 and 2.3.2.

2.3.1.2 Global Statistics and Out-of-Control Criteria

Since the data streams are spatially correlated, we propose using a multivariate CUSUM (MCUSUM) statistic as the global statistic. To be more concrete, the local CUSUM statistics $C_{i,t}$ calculated for those observed sensors at time t are fused into a global CUSUM statistic as follows:

$$GC_t = ||C_{k,t}|| = \sqrt{C_{k,t}\Sigma_{\omega\omega}^{-1}C_{k,t}} \qquad k = \{n: X_n \in \omega\}. \tag{2.9}$$

The process is then deemed to be out-of-control at time t if $GC_t > UCL$, where UCL is a predefined upper control limit. Guidelines to choosing the value of the UCL is discussed in subsection 2.3.2.

Recall that there are two main ways of constructing the MCUSUM statistic as described in (Pignatiello and Runger 1990). The two methods differ in the order in which the accumulation and the quadratic transformation is performed. The first method performs the accumulation first by calculating the individual local CUSUM statistics and then combining them into a single quadratic form. On the other hand, the second method calculates local Hoteling T-square statistic (quadratic form) and then performs the accumulation using a univariate CUSUM on the result. Here we adopt the first approach of MCUSUM in equation (2.9) when constructing the global statistic as it fits well with the framework described in the previous section.

2.3.1.3 Sensor Reassignment

Sensor reassignment is simply reassigning the sensors to the sets of observing sensors $\omega \subset \Omega$ and non-observing sensors $\psi \subset \Psi$ at each time step. Here we propose to

choose the set of observing sensors that maximizes the global statistic in (2.9) as so to have potential to detect the true change quickly. Or equivalently, at each time step before taking any observations, our proposed sensor reassignment method is to choose the set of observed sensor $\omega \subset \Omega$ that is the solution to the following optimization problem:

$$\arg\max_{\omega \subset \Omega} \left\{ C_{k,t} \Sigma_{\omega\omega}^{-1} C_{k,t} \right\}, k = \{n: X_n \in \omega\} \ and \ \omega \subset \Omega \quad (1), \tag{2.10}$$

where $C_{i,t}$ is the local CUSUM statistic of sensor i at time t and Ω is the set of all possible sensor subsets of size q.

While the above optimization problem in (2.10) is well-defined from the mathematical viewpoint, it becomes very challenging to solve it from the computational viewpoint, especially in high dimension situations, as the set of candidate solutions Ω becomes too large. Therefore, below we propose a forward selection greedy heuristic method to solve the combinatorial optimization problem in (2.10).

We start with $\omega = \Phi$; the empty set. The first variable to enter the set ω will be the variable that maximizes equation (2.9) when the cardinality of the set is one. The solution is the variable with the maximum local CUSUM statistic $\{X_i: C_{i,t} \geq C_{j,t} \ for \ all \ j\}$. If we partition the covariance matrix of the standardized data into the following block form,

$$\Sigma = \begin{bmatrix} \Sigma_{\omega\omega} & \Sigma_{\omega\psi} \\ \Sigma_{\psi\omega} & \Sigma_{\psi\psi} \end{bmatrix}.$$

then the inverse can be written as:

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{\omega\omega}^{-1} + \frac{1}{b} FF' & -\frac{1}{b} F \\ -\frac{1}{b} F & \frac{1}{b} \end{bmatrix},$$

where $F = \Sigma_{\omega\omega}^{-1}\Sigma_{\omega\psi}$, and $b = 1 - \Sigma_{\psi\omega}\Sigma_{\omega\omega}^{-1}\Sigma_{\omega\psi}$.Hence if we let $G^{\omega} = C_{\omega,t}\Sigma_{\omega\omega}^{-1}C_{\omega,t}$, the global statistic with respect to the set ω, then the global statistic with respect to the joint set $\{\omega \cup \psi\}$ is

$$G^{\omega\cup\psi} = G^{\omega} + \frac{1}{b}\left\{C_{\psi,t}^{2}(1-F) - C_{\psi,t}F + FF'\right\}, \tag{2.11}$$

which means that the gain in the global statistic after adding variables in set ψ to set ω can be represented by the following:

$$G^{\omega\cup\psi/\omega} = G^{\omega\cup\psi} - G^{\omega} = \frac{1}{b}\left\{C_{\psi,t}^{2}(1-F) - C_{\psi,t}F + FF'\right\}. \tag{2.12}$$

The following variable to enter the set ω will be the variable that maximizes (2.9) when the cardinality of the set is two given that the first chosen variable is X_i. This translates to the variable X_j that maximizes the gain given by (2.12) when the set $\omega = \{X_i\}$ and the set $\psi = \{X_j\}$. Consequently, at any step, the next variable to enter set ω given its current cardinality is the variable that maximizes the gain. The steps at each iteration of this heuristic is illustrated in algorithm 2.1.

Algorithm 2.1: Greedy Forward Sensor Selection to Solve Equation (2.10)	
	Input: Empirical covariance matrix Σ, scalar $r, C_{k,t}$ for all k
	Forward sensor selection strategy: While $(\lvert\omega\rvert < r)$,
1	Calculate the gain $G^{\omega\cup j/j}$ for all variables $X_j \notin$ ω according to eq.(2.12)
2	Augment the set of ω by including $\{X_i: G^{\omega\cup i/i} \geq G^{\omega\cup j/j}\ for\ all\ j\}$
3	Update the global statistic G^{ω}
	End

The initial assignment of sensors in the sets ω and ψ has no significant impact to the monitoring procedure (Liu et al. 2015a). This is due to the adaptive nature of the sampling strategy that reassigns the sensors at each observation time.

2.3.1.4 Overview of the CDS Algorithm

Algorithm 2.2 illustrates the steps of the proposed CDS procedure. Compared to other procedures that assume spatial independence, our approach uses the correlation structure and the information obtained from observed sensors to dynamically compensate unobserved ones. If an unobserved sensor is positively correlated with an out-of-control sensor, then the sensor will be compensated more than the one that is independent. This makes it more likely to choose that sensor in the next step. This property will be demonstrated in the case studies in Section 2.5.

Algorithm 2.2: Correlation based Dynamic Sampling (CDS)	
	Input: Empirical covariance matrix Σ, scalar $r, \delta, \text{UCL}, \alpha \in (0.5,1)$ $C_{k,0} = C_{k,0}^{+} = C_{k,0}^{-} = 0$ for all sensors
	While $(GC_t < UCL)$,
1	Observe sensors based on the current assignments to ω and ψ based on the top-r sensors at time $t-1$
2	For sensor $k \in \omega$, compute the local statistics $C_{k,t}, C_{k,t}^{+}$, and $C_{k,0}^{-}$ according to eq.(2.2)
3	For sensor $k' \in \psi$, compute the local statistics $C_{k',t}, C_{k',t}^{+}$, and $C_{k',0}^{-}$ according to equations (2.3) and (2.4)
4	Reassign sensors to the sets ω and ψ according to Algorithm 2.1 and take observations from the updated set ω
5	Obtain the global statistic GC_t based on the updated set ω from step 4
	End

2.3.1.5 Properties of CDS

This subsection illustrates two behavioral properties of the proposed CDS procedure. These two properties address the desire to disperse sensor deployment when the system is running smoothly under the in-control state, while also quickly localizing at a fault location whenever a true fault occurs. Proofs of the proposed properties can be found in Appendices A.1 and A.2.

Recall that a variable $x_{i,t} \in \omega_t$ if and only if it is observed at time t. Thus, at a given time t_0, the set of sampled variables x_{k,t_0} is given by ω_{t_0}. The following property shows that when the process is in-control or when those variables in ω_{t_0} involve insignificant mean shifts, our proposed sensor deployment procedure will eventually choose variable x_{k',t_0} that does not belong to a neighborhood of ω_{t_0}. This implies the random behavior of our dynamic sampling method under the in-control phase, where sensors will be sampled infinitely many times as the $UCL \to \infty$. This essentially guarantees that the sensor deployment procedure will not permanently localize at any specific location.

Property 1: For a fix time t_0, we assume that $|E[x_k]| \leq \Phi(1-\alpha/2)$ for any $x_k \in \omega_{t_0}$. Consider another variable $x_{k'} \notin \omega_{t_0}$ satisfying $corr(x_{k'}, x_k) = 0$ for all $x_k \in \omega_{t_0}$. Let $UCL \to \infty$, and denote $T_{t,k'} = inf\{t \geq t_0 : x_{k'} \in \omega_t\}$, then $P\left(T_{t,k'} < \infty\right) = 1$.

Next, we will show that when a significant mean shift occurs, our proposed sensor deployment procedure has a greedy property that eventually sticks to the fault area, or to its neighborhood when we do not have enough sensors to cover the whole fault area.

Property 2: Denote the fault area as $\mathcal{O} = \{x_k : |E[x_k]| > \Phi(1-\alpha/2)\}$. Let $UCL \to \infty$, there exists $\mathcal{O}_0 \subseteq \mathcal{O}$ such that $P_1(O_0 \subset \omega_t \; for \; all \; t \geq t_0) = 1$ for some t_0.

In the event that the process is out-of-control, the second property suggests that sensors localized at the fault area will remain deployed within its neighborhood. When a fault is detected in an area, it is desired to check that area as well as its surroundings, because the main issue may be in the neighborhood rather than the initially detected location. Therefore, we are only interested in showing that a remote location, relevant to

the fault area, will not be a point of interest for future sampling. This level of flexibility allows the sampling procedure to better localize around the faulty area rather than simply sticking to an initial suspect area.

2.3.2 Setting Input Parameters

This section will provide guidelines for assigning the values of the input parameters as illustrated in algorithm 2.2. This includes the parameters $\delta, \alpha, r,$ and UCL.

Setting δ: From the literature review presented in Sub-section 2.3.1.1 on the calculation of the local CUSUM statistic, δ represents the smallest mean change magnitude that we are interested in detecting. In practice, the choice of δ can be a target value set using engineering knowledge in the application domain.

Setting r: The choice of r directly affects the detection power of the monitoring procedure. Setting r to be too large will dilute the contribution of the out-of-control sensors to the global monitoring statistic, thereby causing an undesired delay in the detection of the mean shift. Moreover, $r \leq |\omega|$, where $|*|$ denotes the cardinality of a set. The ideal choice for r would be the total number of variables associated with the faults that are of interest for detection, also referred to as the root causes. However, this is usually unknown unless it can be provided from engineering knowledge. In the case that it is unknown, choosing a small value of r has been shown to be robust to various fault types (Mei 2010).

Setting UCL: The UCL is the threshold that determines when to stop the monitoring procedure and alert the detection of a change. The value of UCL is related to the pre-scribed in-control ARL of the monitoring scheme. The practitioner can determine the optimal UCL

value from sufficiently large in-control measurements or via Monte Carlo simulation and bootstrap techniques (Efron and Tibshirani 1994, Chatterjee and Qiu 2009).

Setting α: The tuning parameter $\alpha \in (0,1)$ is a very crucial parameter that essentially determines the trade-off between how sporadic the behavior of the algorithm is when the process is in-control and how fast it converges to the faulty sensors when the process is out-of-control. To illustrate this further, as α approaches 0, the local statistic compensation provided to variable $k \in \psi$ will exceed that of variable $k \in \omega$. While this is not an issue when the process is in-control, the algorithm will not be able to converge to a unique set ω when the process goes out-of-control, as there will always be a variable in ψ with a larger local statistic. On the other extreme, if α approaches 1, variables belonging to set ψ will receive almost no compensation causing the sensor assignment of the algorithm to be static which is clearly undesired.

To narrow down the choice of α we can initially try to find tighter bounds. From the proof of properties 1 and 2 of our CDS algorithm, the compensation requires $\delta < \Phi^{-1}(1-\alpha/2) < |\delta^*|$. Here, δ^* is the true mean when the process goes out of control. Generally speaking, δ^* is unknown and this makes it challenging to get an upper bound.

In order to obtain an appropriate value for α, we simulate the monitoring procedure iteratively with a binary search over the range of α. The criteria for terminating the search is when the percent decrease in standard deviation (denoted by υ) of the number of times (denoted by η) that each variable is assigned to set ω is less than some predefined value ζ. The details of this procedure are outlined in algorithm 2.3. The intuition is to determine a

choice of α that provides minimum deviation between sensor sampling frequencies while maintaining the pre-specified ARL.

Algorithm 2.3: Choosing the value of α	
	for $t = 0$, set $\alpha_t = \begin{cases} 2[1-\Phi(\lvert\delta^*\rvert)], & \text{if } \delta^* \text{ is known} \\ 2[1-\Phi(2\delta)], & \text{o.w} \end{cases}$ $v_t = \mathrm{M}$, where M is sufficiently large
	for $(i = 1{:}I)$, (I: Maximum number of iterations)
1	Generate N instances of n in-control observation for all sensors
2	Run algorithm (1) for each instance j calculating $\eta_{t,j}$ and $v_{t,j} = Var[\eta_{t,j}]$
3	Let $v_t = E[v_{t,j}]$
4	If $\lvert v_t - v_{t-1}\rvert / v_{t-1} < \zeta$; break loop
5	set $\alpha_t = \begin{cases} \alpha_t/2, & v_t < v_{t-1} \\ 3\alpha_t/2, & o.w \end{cases}$
	End

2.3.3 Estimating the Precision Matrix

To effectively implement our proposed CDS algorithm, it is important for us to be able to obtain a reasonable estimation of the inverse covariance matrix in a high-dimensional setting, also known as the precision or concentration matrix (Hsieh et al. 2011). In practice, the precision or inverse covariance matrix can be either learned from historical training data or imposed by the domain knowledge, and the following two subsections will discuss these two approaches, respectively.

2.3.3.1 Covariance Structure Estimation

The estimation of the precision matrix in high dimension from training data has been an area of interest for many researchers in the past years, since it can provide information on the interrelations between variables in graphical models (Scheinberg et al. 2010).

A sparse representation of the inverse covariance matrix is desired in the estimation process in high dimensional settings, due to the advantages that sparsity offers. When the number of observations is limited, as is the case in many modern high-dimensional statistical problems, sparsity promotes robustness to the estimation process which translates well to the future observations (Duchi et al. 2012). Moreover, inducing sparsity functions regularize and enhance interpretability and counter overfitting (Scheinberg et al. 2010).

Methods to estimate the precision matrix look into solving the following optimization problem, its dual or some variation of it:

$$\log\det\theta - tr(S\theta) - \rho|\theta|_1, \tag{2.13}$$

where $\theta = \Sigma^{-1}$, and S is the empirical covariance matrix.

The objective function in (2.13) is a convex problem that can be solved with interior point methods in $O\left(p^6 \log(1/\varepsilon)\right)$, however this becomes infeasible for even moderate p. Banerjee et al. (2008) used block coordinate decent with a cost of $O(p^4)$ with their proposed algorithm COVSEL. By solving iterative LASSO problems, the graphical LASSO algorithm proposed by Friedman et al. (2008) manages to reduce the computation complexity to $O(p^3)$. The greedy gradient ascent method and alternating linearization methods (Scheinberg and Rish 2009, Scheinberg et al. 2010), as well as the projected subgradient method developed by Duchi et al. (2012) all claim to reduce the complexity to $O(p^2)$. The second order algorithm QUIC proposed by Hsieh et al. (2011) solves iterative quadratic approximations that has a reduced cost of $O(p)$ to find a Newton direction.

2.3.3.2 Covariance Structure Imposition

There are several domains where prior knowledge of the system being monitored can be used to extract some process characteristics that can help bypass the estimation of the inverse covariance matrix by alternatively imposing a covariance structure on the data streams. A prominent example of such an application is when the data is acquired in the form of images.

If we regard each pixel of an image to be a variable for monitoring purposes, then we can assume that the value of any pixel is independent of other pixels given its neighborhood. This assumption can be translated to a special structure by imposing that the entries in the precision matrix that correspond to two pixels that are not within a certain pre-specified proximity is equal to zero. This level of proximity represents the closeness of the values of nearby pixels. Naturally, different areas of an image can have a different level that is suited to the correlation of the pixels in set area. This type of structure imposition will be demonstrated in the solar flare case study discussed in sub-section 2.5.1.

2.4 Simulations

This section serves as an evaluation for the performance of our proposed CDS algorithm. We compare the performance to two state of the art algorithms, TRAS (Liu et al. 2015a) and Top-r (Mei 2010). It is very important to note that the Top-r method assumes no limitations in the number of variables that can be observed and thereby has full access to all raw sensors or data streams. We include it in the comparison to illustrate how competitive our proposed method is, even when compared to those without sampling limitations.

In our simulations, the data is generated using the following generative model:

$$X_t = AZ_t + \varepsilon_t, \tag{2.14}$$

where the observed variables at time t are $X_t \in R^p$, latent variables $Z_t \in R^q$ following a multivariate normal distribution $MN(0, I)$, and white noise $\varepsilon_t \in R^p$ following $N(0, \sigma_\varepsilon I)$. Matrix $A \in R^{p \times q}$ that maps the latent variables into the domain of the observed variables. Hence, the observed variables follow a multivariate normal distribution as well with $MN(0, AA^T + \sigma_\varepsilon I)$.

In the generative model described above, the transformation matrix A controls the sparsity in the covariance of the observed variables X_t. If the matrix A is block diagonal, such that each block is of size $p_i \times q_i$ with $\sum_i p_i = p$ and $\sum_i q_i = q$, then the covariance matrix of the observed variables X_t will also be block diagonal with blocks of sizes $p_i \times p_i$. Therefore, as we decrease the block size in the transformation matrix A, we induce a higher level of sparsity in the observed variables X_t. In our simulations, we chose $p = 1500$ and $q = 150$. The blocks in the transformation matrix are of size $p_i \times q_i = 100 \times 10$ for all i, and each block is a random matrix whose entries are i.i.d. Uniform(-1,1) random variables. The control limits were chosen to achieve an in-control ARL of 200. The shift is introduced in a single block of latent variables, however only 150 variables from the full observations X_t could be obtained at any given time (i.e $m = 150$). Out of the 150 available observations, the test statistics are constructed using $r = 15$ variables.

Table 2.1 demonstrates that the CDS algorithm consistently outperforms the TRAS algorithm by an average 42% reduction in detection delay. Moreover, it is interesting to compare our proposed method to the Top-r procedure, which assumes no limitations on

data acquisition. Although it may be expected that it would be better than our proposed CDS procedure due to the full visibility, the detection delay of CDS within 3.5% from the Top-r and can even surpasses it. This can be attributed to the fact that the global monitoring statistic of our proposed CDS method takes into account the correlation of the data streams rather than the independence assumption of the other two competing methods.

Table 2.1 Performance evaluations of the CDS algorithm under different shift magnitudes as well as comparisons with Top-r and TRAS

	In-control ARL (standard deviation)			**Out-of-control ARL (standard deviation)**		
Shift size	**Top-r**	**TRAS**	**CDS**	**Top-r**	**TRAS**	**CDS**
$\boldsymbol{\delta = 0.25}$	214(210)	222(186)	220(182)	56(23)	74(42)	*51(31)*
$\boldsymbol{\delta = 0.5}$	212(189)	226(172)	223(180)	*24(14)*	50(31)	34(21)
$\boldsymbol{\delta = 1}$	201(175)	205(181)	210(172)	*11(5)*	17(8)	12(7)
$\boldsymbol{\delta = 2}$	207(182)	210(188)	197(210)	8(2)	8(3)	*5(2)*
$\boldsymbol{\delta = 4}$	221(193)	220(213)	210(185)	1	1	1

2.5 Case Studies

This section presents a study on two real datasets to showcase the capability of our adaptive monitoring procedure in practical scenarios. The first subsection 2.5.1 illustrates how the correlation based adaptive method can achieve high performance under limited transmission capacity by leveraging partial images obtained from video recording of solar flare occurrences. The second subsection 2.5.2 demonstrates how adaptive sampling can be utilized to monitor in-line Raman spectroscopy for CNTs manufacturing.

2.5.1 Solar Flare Detection

The detection of solar flares via satellite imaging is an example of a monitoring process that generates high dimensional data, where the occurrence of solar flares is regarded as the change (defect). The solar flare phenomenon consists of various dynamical processes that take place in the solar atmosphere, where a resulting visible brightening of the emission constitutes a flare that can last from 1-15 minutes (Parker 1963). The energy released from this phenomenon can interfere with radio communications by disturbing Earth's ionosphere (Augusto et al. 2011). This serves as motivation to detect these flares upon onset with minimal delay. At each second during the satellite's recording, tons of solar flare images will be captured and generated. The imaging instruments typically have memory capacities of 16TB and can acquire images at a rate of 25 frames per seconds. Due to the enormous amount of data and limited memory of the imaging instruments, only one data set every second can be archived (Ishii et al. 2013). However, due to the transient characteristic of the solar flare process and large amount of dataset, methodologies for image change detection such as (Yan et al. 2018) that analyze the full data stream can usually exceed the transmission and processing capabilities during online monitoring, and thus may not be suitable for detecting solar flares in real time.

The solar flare dataset used in this study is publically accessible in video format at http://nislab.ee.duke.edu/MOUSSE/index.html. The data is collected from satellite images that are taken in very high frequencies. At each frame 232×292=67744 of high dimensional data are observed and there are 300 frames in the video. This is a very high dimensional dataset to process especially when the number of available observations is very small in

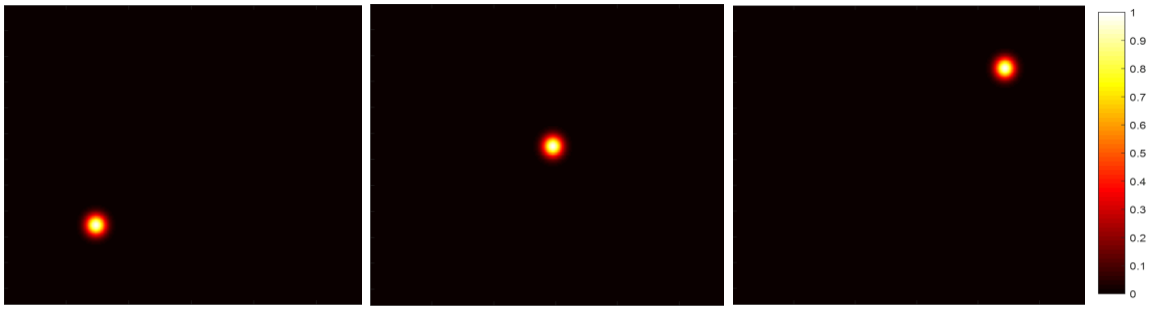
relation. There are two clear occurrences of solar flares that are visible at frames t=187~202 and t=216~268, respectively.

Some preprocessing has to be done to the raw data before implementing the proposed methodology. In order to rescale the original pixel intensities obtained from the raw video, we perform background removal by differencing the data with a moving average window of size 4. This is intended to remove the autocorrelation between successive frames as well as to normalize the pixels. To put it in mathematical terms, the processed data X'_t is calculated through the relation:

$$X'_t = X_t - \frac{1}{4}\sum_{i=1}^{4} X_{t-i}, \quad \text{for } t \geq 5$$

$$X'_t = X_t - \frac{1}{t-1}\sum_{i=1}^{t} X_{t-i}, \text{ for } 2 \leq t \leq 4 \text{ , and } X_1 = 0.$$

The remaining data after removing the background was found to be approximately normal, as was the case in the study in (Xie et al. 2013).



(a) (b) (c)

Figure 2.1(a), (b) and (c) illustrate the imposed covariance structure over the three different pixels.

In this particular study, the parameters are selected as $\delta = 1$, $\alpha = 0.27$, which corresponds to $\Delta = 0.1$. It should be noted that several manipulations of the previous

parameters also yield similar results to the ones illustrated here. We further assume that the number of pixels that can be transmitted for analysis at any acquisition time is 1000 out of the available 67744 pixels in a full frame (image) of the video. In other words, the parameter q is equal to 1000, while we set $r = 40$.

As for the spatial covariance structure, we choose to impose the precision or inverse covariance matrix as in subsection 3.3.2 to be exponentially decaying with radius of 20 pixels. This is generally appropriate for images, particularly for the solar flare which often occurs in a local area. This specific covariance imposition is demonstrated by Figure 2.1, where three plots illustrate the imposed covariance structure over the three different pixels (17107, 34214, 51321). For example, Figure 2.1(b) is an image that is obtained when the $34{,}214^{\text{th}}$ row/column vector of the 67744×67744 pixels covariance matrix is reshaped into a 232×292 matrix, which corresponds to the dimensions of a video frame. This serves to illustrate that any given pixel is only correlated with other pixels in its proximity.

In Figure 2.2 and Figure 2.3, the images in (a) shows original frames of the captured video, while (b) and (c) illustrate the sampling (pixels in white) using TRAS and our proposed methodology (CDS), respectively. Frame 186 is approximately the frame that precedes the first solar flare occurrence. The figure shows that both methods behave in a random fashion, which is desirable since the process is essentially still in control (i.e. a flare has yet to occur). This can also be seen from the images at frame 215, before the second flare, which also serves to demonstrate the capability of our CDS algorithm to return to the random behavior after the end of the first flare.

Frame 198 represents the moment when the solar flare is the brightest. Figure 2.3 illustrates the sampled pixels at this frame as well as frame 230, when the second flare is brightest. Our proposed CDS algorithm covers the flare area completely in both occasions. On the other hand, they are only partially covered when using methods that do not consider the correlation structure.

In addition to the ability of our proposed CDS algorithm to localize at flare location, Figure 2.4 demonstrates its capability to localize faster than the competing TRAS algorithm. Figure 2.4 (a),(b) show the sampled pixels right before and right after detection by the competing TRAS algorithm, at frames 194 and 195 respectively. While, Figure 2.4 (c),(d) show the sampled pixels right before and after detection by the CDS algorithm, at frames 190 and 191 respectively. The ability of the CDS algorithm to outperform the TRAS algorithm, with regards to detection delay, can be attributed to the significantly faster localization. This can be vividly observed from the instantaneous localization within a single frame.

With only 1.5% pixels available from the 67744 pixels per frame, our proposed algorithm can detect the flare at frame 191; only 4 frames after its onset at frame 187. Liu et al. (2015a) reported the detection of the change at frame 190 when 2000 pixels were observed at any time. While as shown in the figure, this performance deteriorates to frame 195 when the amount of pixels is cut to 1000. While, our proposed CDS algorithm with only half of the resources, can still compete with that performance, due to the superior localization strategies.

Figure 2.5 plots the global monitoring statistic of the proposed CDS algorithm from frame 100 to the end of the captured video at frame 300. For comparison, Figure 2.5 (b) illustrates the monitoring statistic obtained from the competing TRAS algorithm. The first 100 frames were considered a training sample and were used to obtain the control limits using a bootstrap procedure. The control limits for both CDS and TRAS algorithms were set to a pre-specified in-control ARL of 2500 were determined to be 970 and 950, respectively. The occurrence of the second flare was very close to the first and therefore Figure 2.5 only shows the monitoring statistic crossing the threshold once. This is because the 14 frame difference between the end of the first flare and the beginning of the second is insufficient to reset the declining statistic. In such scenarios, the statistic can be simply reset upon resolving the preceding out of control occurrence. In this study, the monitoring statistic was reset at frame 203 after the end of the first flare.

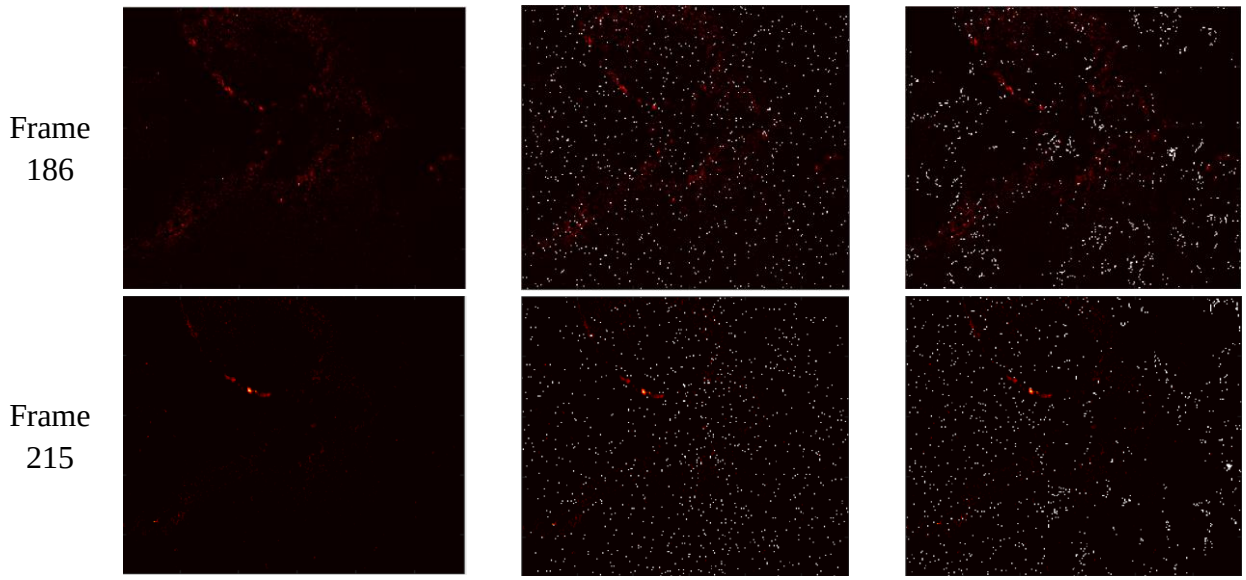


Figure 2.2 Monitoring frames before the two flares: (a) frame capture from video; (b) sampled pixels from the TRAS algorithm; (c) sampled pixels from the proposed CDS algorithm

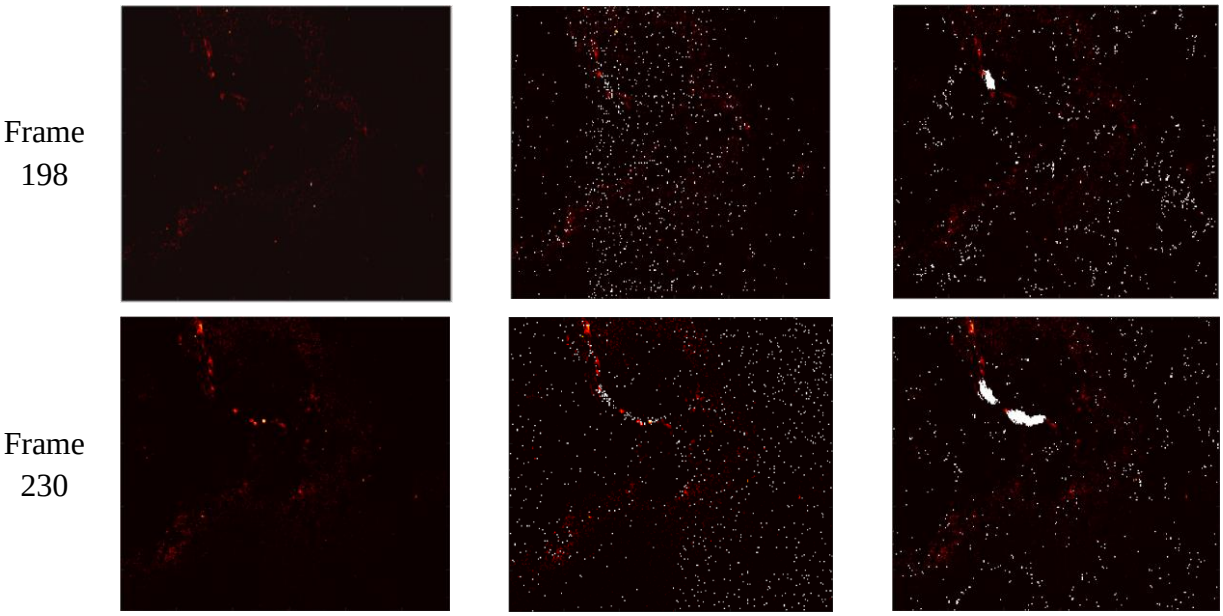


Figure 2.3 Monitoring at the solar flares peak: (a) frame capture from video when the flare is the brightest; (b) sampling from TRAS; (c) sampling from CDS.

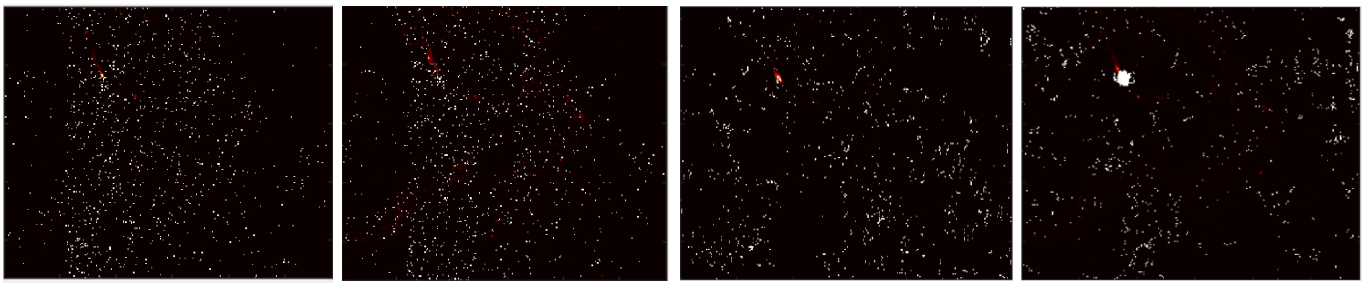
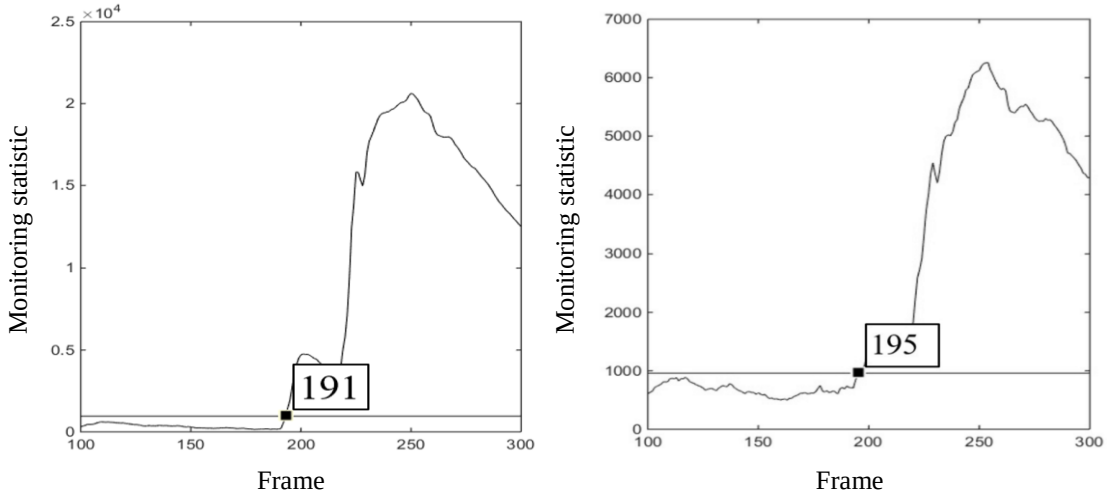


Figure 2.4 Detection of the first solar flare: (a), (b) sampling from TRAS right before and after detection; (c), (d) sampling from CDS right before and after detection.

Similarly, the CDS algorithm is capable of detecting the second flare at frame 219, only 3 frames upon onset. While, the competing TRAS algorithm lags behind by 7 frames, and detects the flare at frame 223. The detection in 3 frames not only beats the TRAS

algorithm under the same limitations, but also outperforms the reported detection time of frame 221 reported in (Liu et al. 2015a), which had double the visibility.



(a) Proposed CDS algorithm performance

(b) TRAS algorithm performance

Figure 2.5 The monitoring statistics by respectively implementing the CDS/TRAS algorithms with the detection frames illustrated by the data cursors.

2.5.2 Fault Detection for in-Line Raman Spectroscopy

In this subsection we evaluate the performance of our methodology in addressing the challenges of monitoring the production process of continuous carbon nanotubes (CNTs) buckypaper using inline Raman spectroscopy. We aim to show that settling for adaptively sampled partial signals to improve acquired signal quality can be a worthwhile tradeoff.

The monitoring of the manufacturing process of CNTs buckypaper manufacturing in real time using in-line Raman spectroscopy has gained much interest recently (Yue et al. 2018). The ability to monitor this process in real time is critical to scale it up while meeting high quality standards. However, it is challenging to detect changes in the data

collected from this procedure since there are several sources for variation in Raman spectrums. One source of variation is the acquisition frequency of signals (Yue et al. 2017b). Characterization of an inline Raman spectrum may take multiple scans with a sampling frequency of ten seconds to several minutes. Higher frequencies result in higher signal to noise (S/N) ratios due the rapidly moving samples. Figure 2.6 illustrates acquired Raman spectrums from two operating conditions (red, blue), where both the acquisition frequency and signal intensity are higher in the second operation settings. In this case, it may be beneficial in terms of detection time to acquire partial signals at a higher frequency, thereby maintaining the same S/N ratio and signal quality as the lower acquisition frequency.

This study will compare our CDS procedure against the same benchmark methods in the other studies; TRAS and Top-r. The Top-r method requires full observations and therefore will be applied to data with low S/N ratios. While the two adaptive monitoring schemes (CDS, TRAS) will be implemented on partial data with high-signal to noise ratio as illustrated by the red profiles of Figure 2.6 (left). The data set consists of 200 in-control profiles and 50 out of control instances, where the dimension of each profile is $p = 512$. In order to obtain signals with high S/N ratios at the same high frequency, approximately 10% of the Raman spectra ($m = 50$) can be measured at any given time. For each method, a threshold that satisfies an in-control average run length (ARL) of 500 is determined by bootstrapping the 200 in-control samples. The remaining parameters are set to the following: $r = 25, \delta = 1$ and the compensation significance level was found to be $\alpha = 0.23$ ($\Delta = 0.21$) via algorithm 2.3. Figure 2.6 illustrates the mean of out of control signals,

where it can be noted that the shift approximately within the index interval [95,115] of the Raman spectrum.

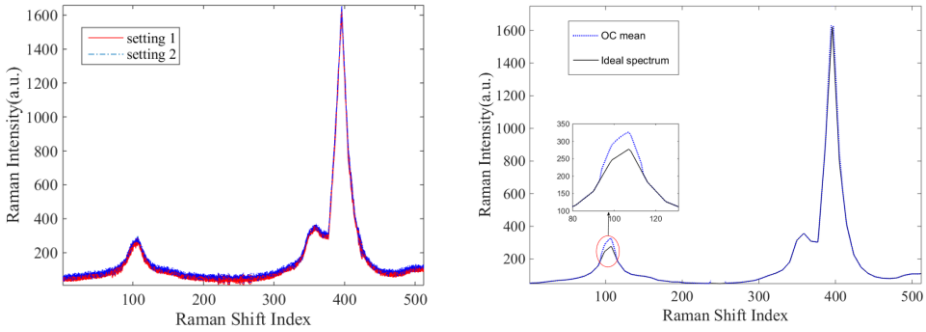


Figure 2.6 Left: illustration of the Raman spectra data. Right: illustration of out-of-control Raman spectrum mean shift.

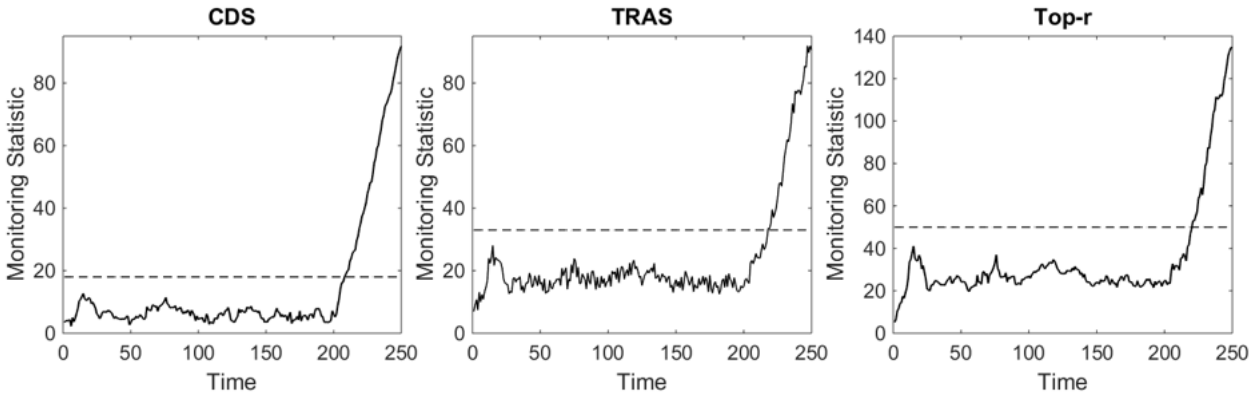


Figure 2.7 Monitoring statistics for in-line Raman spectra

Finally, the covariance matrix is estimated from the first 100 in-control data using the method QUIC (Hsieh et al. 2011), which is the precision matrix estimation technique discussed in sub-section 2.3.3.1.

The results from implementing the different monitoring schemes are presented in Figure 2.7. Our CDS procedure outperforms the other benchmark methods and signals an alarm at time 209, which is 9 epochs upon failure onset. Since TRAS does not take into

account the correlation structure between variables, it is unable to quickly localize at the fault region. This results in a detection delay of 18 epochs. Finally, the Top-r procedure achieves the lowest performance among the three with a detection delay of 21 epochs, even though it was implemented on complete data. This is because complete data collected at a high frequency comes at the cost of low S/N ratios as we have discussed in the introduction of this study. Hence, we can conclude from this study that there are practical scenarios where it may be beneficial to sacrifice sensor visibility in exchange for better quality data. This emphasizes the importance of making educated decisions on which sensors to acquire in real time.

2.6 Conclusion

The development in sensing technology that generate high dimensional data has offered unprecedented process monitoring capabilities. However, with this advancement rose new challenges that require novel monitoring schemes in limited resources due to sensors availability for deployment, transmission capacity and computational power. Hence, it is useful to apply the multi-armed bandit algorithms to the SPC context to tackle the issue of efficient monitoring under the limited resources environments.

This chapter proposes a novel correlation based dynamic sampling strategy that constructs a dynamic compensation factor to unobserved data streams. This is performed by using the idea of celebrated upper confidence bound (UAB) algorithm from the multi-armed bandit problem, as well as by utilizing the correlation structure between the observed and unobserved streams. A novel integration of the top-r procedure with multivariate CUSUM is developed to construct the global monitoring statistic used for decision making

related to the state of the process. This results in a strategy that is effective in monitoring high dimensional data streams with partial observations, which consequently reduces the computational cost at the data fusion center. Moreover, utilizing the correlation structure embedded in the data streams allows for faster localization at the fault source while maintaining a random sampling behavior when the process is in-control, which was illustrated by the two properties of the dynamic sampling behavior. This allows this method to be suited for a wide area of applications, such as network processes and images as was demonstrated in the solar flare case study. In addition to advanced industrial manufacturing operations as showcased by the in-line Raman spectroscopy case study.

CHAPTER 3. HIGH DIMENSIONAL PROCESS MONITORING USING ROBUST SPARSE PROBABILISTIC PRINCIPAL COMPONENT ANALYSIS

3.1 Introduction

With the deployment of large numbers of sensors and the wide use of imagery in monitoring, the monitoring of high dimensional data streams that result from these systems has gained a lot of interest in recent years. Traditional statistical process monitoring procedures may fall short in these data rich environments. A popular approach to address this issue is to reduce the dimension of the available data streams. Principal Component Analysis (PCA) is a ubiquitously used dimension reduction technique (Jolliffe 2011). PCA is a method that projects a set of observed variables onto a significantly lower dimensional subspace spanned by directions referred to as principal components. The resulting projection points are commonly referred to as PC scores. However, PCA has been shown to produce extremely inconsistent estimates in high dimensional settings, when the low dimensional space is sparse (Ma 2013). Not to mention the inherent issues that arise from poor interpretability as the estimated principal components are linear combinations of all the data streams (Archambeau and Bach 2009, Guan and Dy 2009).

An example of such data rich environments is the continuous production of carbon nanotubes (CNTs) buckypaper. A recent development in the inspection process utilizes in-line Raman spectroscopy (Yue et al. 2018). The ability to monitor the manufacturing process in real time is critical to scale it up while meeting high quality standards. However, it is challenging to detect changes in the data collected from this procedure. This is because the obtained profiles are high dimensional with specific segments where peaks occur as

illustrated in Figure 3.1. In addition to the sparsity of these features, the noise is complex with signal-dependent properties and may be confused with defects (Yue et al. 2017a).

In this chapter, we propose a new method that combines the two properties of sparsity and robustness within a probabilistic framework to facilitate the monitoring of Buckypaper production process. Preceding work on the problem of obtaining both sparse and robust principal components include (Hubert et al. 2016, Croux et al. 2013), which will be used as benchmarks in the principal component extraction portion of the simulated experiments. The aforementioned work combine the two properties by developing sparse modifications of existing robust formulations of PCA, namely projection pursuit PCA (PP-PCA) (Croux and Ruiz-Gazen 2005, Li and Chen 1985) and ROBPCA (Hubert et al. 2005).

While the aforementioned methods may be effective for the objective of extracting robust and sparse principal components, their implementation for statistical decision making during process monitoring can be complicated. This is because the extraction is not achieved in a probabilistic space and uniform distance measures (e.g. the Euclidean distance) are used. However, all statistical conclusions are based in a probabilistic space where a Mahalanobis type distance measure is more appropriate (Kim and Lee 2003). Therefore, we develop a sparse and robust method using a probabilistic model for the purpose of implementation in process monitoring. A probabilistic approach provides a direct platform for change detection and fault diagnosis, which will be used to address the monitoring challenges of the motivating Buckypaper production process. In addition, a probabilistic method has potential in addressing incomplete or missing data via conditional probability densities as well as extensions to mixture models (Archambeau et al. 2008, Tipping and Bishop 1999a). However, these properties will not be discussed in this study as they deserve to be autonomously investigated.

In the case study and the monitoring portion of the simulation experiments, we evaluate the performance of our robust and sparse probabilistic approach for process monitoring by comparing it with other probabilistic methods that lack one or both properties. Several probabilistic PCA variations have been proposed for process monitoring including, some robust and some sparse but to the best of our knowledge none that combine both (Zeng et al. 2017, Zhu et al. 2014, Chen et al. 2009).

The remainder of this chapter is organized as follows: In Section 3.2, we provide a brief review of relevant topics in the literature followed by a more detailed overview of dimension reduction methods dealing with sparsity and/or robustness. Next, in Section 3.3, we illustrate in detail our proposed dimension reduction methodology as well as the monitoring strategy. Section 3.4 demonstrates the effectiveness of our proposed sampling strategy in virtually simulated scenarios. Furthermore, section 3.5 presents a case study on change detection for in-line Raman spectroscopy. We then finally conclude the chapter with a discussion of the major findings of our proposed monitoring scheme.

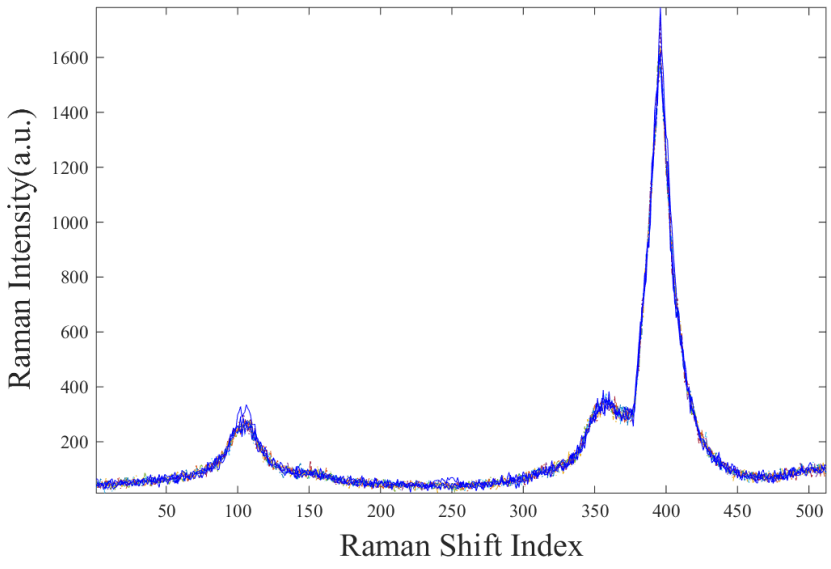


Figure 3.1 Illustration of the Raman spectra data

3.2 Literature Review

This section is split into two subsections. The first subsection 3.2.1 provides a review of classical PCA and various methodologies used to improve robustness and induce sparsity to the original formulation of PCA. While the second subsection 3.2.2 presents a brief overview of the basic approach for probabilistic principal component analysis (PPCA) and its robust and sparse variations. This progression will lay the necessary foundation necessary to facilitate the discussion of our proposed robust sparse probabilistic PCA method in section 3.3.

3.2.1 Classical, robust and sparse PCA

Classical PCA is a dimension reduction technique that projects a set of observed variables $x \in \mathbb{R}^p$ onto a subspace of latent variables $z \in \mathbb{R}^q$, where the dimension of the latent variables is significantly lower than the dimension of the observed variables; i.e. $p \gg q$. The objective is thus to find the principal components, which are a linear combinations forming what is known as the loading matrix denoted by $A \in \mathbb{R}^{p \times q}$. This is achieved by searching for the directions that result in PC scores with maximum variances. In doing so, the resulting principal components and their variances correspond to the eigenvectors and eigenvalues of the covariance matrix Σ of the observations x.

Traditional PCA deals with the L_2 norm, which is optimal when we are concerned with minimizing the Mean Square Error (MSE) (Wang et al. 1996). However, when PCA is applied in noisier environments with significant outliers, it becomes increasingly important for a more robust measure. In some cases, outliers can be removed from the estimation process, making regular PCA adequate. However, in practice we do not know

which data points are outliers, especially in a high dimensional setting. Not to mention that in such a setting, each data point is too valuable to be discarded. This is due to the usual scarcity of observations relative to the dimension. In the literature, the problem of modeling data sets with erroneous entries and outliers while simultaneously detecting them is referred to as the robust PCA problem.

Iterative Reweighted Least Squares (IRLS) is a straight forward algorithm for obtaining robust components (De La Torre and Black 2003). The basic idea is to iteratively apply regular PCA while down-weighting poorly fitted observations between iterations. Another approach proposed by Candès et al. (2011) assumes that the data is a superposition of a low rank and a sparse component. The objective is then to find the decomposition that minimizes a weighted mixture of the nuclear norm and the L_1 norm. This method solves a convex program called the Principal Component Pursuit (PCP). A detailed review of these algorithms as well as other variations and extensions can be found in (Vidal et al. 2016).

Classical and robust variations of PCA obtain a lower dimensional subspace that is spanned by a linear combination of all the variables in the original high dimensional subspace. This results in interpretability issues especially in data rich environments. This shortcoming of PCA can be addressed by adjusting the original formulation such that a relatively small number of nonzero entries are allowed for the loadings. These nonzero elements will thus correspond to the features that contribute the most information in the data population. Such formulations are commonly referred to as Sparse Principal Component Analysis (SPCA) (Zou et al. 2006).

Several methods for obtaining these sparse principal components have been explored in the literature. One intuitive approach that was proposed by Cadima and Jolliffe (1995), is to threshold loadings with a small absolute value to zero. This is generally referred to as "simple thresholding" (d'Aspremont et al. 2008). The Simplified Component Technique-LASSO (SCoTLASS) introduces a bound on the sum of the loadings, thereby forcing some of them to become zero (Jolliffe et al. 2003). The sparse low-rank approximation (SLRA) algorithm proposed in (Zhang et al. 2002, 2004) computes matrix low-rank approximations with sparse factors, which is then formulated as a penalized optimization problem. Another SPCA algorithm, introduced by Zou et al. (2006), reformulates the traditional PCA problem as a regression problem. Then, it adds a LASSO (Tibshirani 1996) type penalty, which is a penalized regression technique based on the L_1 norm. Both previous formulations result in non-convex optimization problems that can cause computational issues. d'Aspremont et al. (2008) later introduced another approach that directly incorporates a sparsity condition in the SPCA problem formulation, which resulted in a convex relaxation of the original problem. More recently, Ma (2013) proposed a new iterative thresholding approach. Under a spiked covariance model, this approach was shown to obtain the leading principal components more consistently in sparse high-dimensional settings.

High dimensional data often contain outliers as well as sparse data structures. Few work has been done to combine the two properties of robust and sparse principal component analysis. Most notably Robust Sparse Principal Component Analysis (RSPCA) and (ROSPCA) introduced in (Hubert et al. 2016, Croux et al. 2013). The former combines the two properties by applying the L_1 penalty to the projection pursuit (PP) approach for

obtaining robust principal components. While the latter incorporates sparse PCA within the framework of the robust method ROBPCA, by first finding a robust subspace and then uses the sparse method SCoTLASS to yield a sparse loading matrix. These two approaches for finding both sparse and robust subspaces will serve as a benchmark in the simulation study of section 3.4.

3.2.2 Probabilistic PCA

The classical formulation of principal component analysis is not a probabilistic model in the sense that the latent variable z is viewed as a projection of the observed variables x onto a linear correlation subspace of interest. Therefore, it emphasizes more the observed variables x rather than the latent variables z. Probabilistic principal component analysis uses a probabilistic generative model that was introduced by Tipping and Bishop (1999b). It is called a probabilistic generative model because it models the observed variables as if they are generated from the latent variables with some Gaussian error, while assigning probability distributions to them. This puts the latent variables z in the forefront while the observed variables x are treated as a byproduct that results from a linear combination of the latent variables z.

Based on this probabilistic interpretation of PCA, other variations using Bayesian variational inference to improve robustness have been proposed in the literature (Archambeau et al. 2006, Gao 2008). A latent variable view of the Student-t distribution is used by Archambeau and Bach (2009) instead of choosing a Gaussian noise for the error. While, Gao (2008) replaces the conventional Gaussian distribution for the noise by the Laplacian distribution, also referred to as the L_1 distribution. Both distributions are

characterized by their heavy tails which allows them to be more robust to outliers as opposed to the traditional Gaussian distribution. Moreover, both the Student-t distribution and the Laplacian distribution can be written as a superposition of an infinite number of Gaussian densities. The use of these alternative priors will be detailed further in the methodology part of this chapter in section 3.3.

In addition to the formulations mentioned above, a different probabilistic implementation that focuses on decomposing the latent variable into a low-rank component and a sparse component can be found in (Han et al. 2017, Ding et al. 2011). While both studies propose methods for solving this decomposition, Han et al. (2017) assumes that the sparse component is structured rather than consisting of independent variables.

Variants of PCA probabilistic formulations, that promote sparsity, have also been discussed in the literature. Guan and Dy (2009) proposed using three different sparsity inducing priors (Laplacian, inverse-Gaussian and Jeffrey's prior) for the loading matrix in a Bayesian probabilistic formulation of PCA. In a more general framework, a probabilistic projection model was introduced in (Archambeau and Bach 2009), with an application to sparse PCA as a special case. The use of these alternative priors will be detailed further in the methodology part of this chapter in section 3.3.

3.3 RS-PCA Methodology

The following is a description of the problem we address in this chapter. Suppose we are observing a process with p variables. The observed data at time t is represented by the p dimensional vector $\boldsymbol{x_t} = (x_{1,t}, \dots, x_{p,t})'$. Our objective is to monitor the data acquired and detect any shifts in any components of the data streams. In our analysis we assume that

the data can be projected on a lower dimensional subspace spanned by a set of latent variables $\boldsymbol{z_t} = (z_{1,t}, \dots, z_{q,t})'$. If a sparse representation of the observed variables is desired, then it is reasonable to assume that the underlying spatial structure is given by a block diagonal covariance matrix Σ of data streams. In other words, data streams fall into K disjoint sets $B = \{1, \dots, K\}$ such that $\text{cov}(x_{i,t}, x_{j,t}) = 0$ if $X_{i,t}$ and $X_{j,t}$ belong to two different sets. This assumption is directly related to the interpretability of the model. Thereby, each estimated principal component can only be a linear combination of variables belonging to a single set. Hence, the sparsity of the components dictates the interpretability of the model. The objective of the study is to consistently identify these sparse components in data rich environments with outliers.

The assumption for the noise of the observed data is the Gaussian distribution in the conventional probabilistic PCA as discussed in subsection 3.2.2. The popularity of the Gaussian distribution can be attributed to the attractiveness of the central limit theorem. In many cases, it is justified by the knowledge of the process. Nevertheless, it is usually chosen because of the appealing analytical properties. This assumption may be very limiting in practice when the noise does not follow a Gaussian distribution.

Starting with the basic Gaussian prior for the error ε, we get the following generative model:

$$x = m + A\, z + \varepsilon, \tag{3.1}$$

$$z \in \mathbb{R}^q \sim \mathcal{N}(0\,, \mathbf{\Phi^{-1}}),\ \varepsilon \in \mathbb{R}^p \sim \mathcal{N}(0\,, \sigma_\varepsilon^2 \cdot \mathbf{I}) \text{ and } x \in \mathbb{R}^p \sim \mathcal{N}(0, A \cdot A^T + \sigma_\varepsilon^2 \cdot \mathbf{I}).$$

Here, m is the mean and $A \in \mathbb{R}^{p \times q}$ is the loading matrix. Given the observed data stream x, the parameter set $\{\Theta = \Phi, A, \sigma_\varepsilon^2\}$ can be estimated using an expectation maximization algorithm. For the remainder of this chapter, we assume that the mean is $m = 0$.

The generative probabilistic formulation that induces sparsity and provides robustness is similar to the one described by equation (3.1) except some distributions are replaced with Laplacian priors. The next subsections will demonstrate how this chapter adjusts the basic probabilistic generative model to attain the robust and sparse properties.

3.3.1 Probabilistic Model Reformulation

The probabilistic generative model (3.1) is reformulated to induce sparsity and promote robustness by assuming the Laplacian prior on the latent variables and the noise as follows:

$$p(A_{i,j}|\lambda) = \sqrt{\frac{1}{2\lambda}} \exp\left\{-\sqrt{\frac{2}{\lambda}}|A_{i,j}|\right\}, \qquad i \in \{1, \dots, p\}, j \in \{1, \dots, q\}, \tag{3.2}$$

$$p(\varepsilon_i|\gamma^{-1} = \sigma_\varepsilon^2) = \sqrt{\frac{\gamma}{2}}\ \exp\{-\sqrt{2\gamma}|\varepsilon_i|\}, \qquad i \in \{1, \dots, p\}, \tag{3.3}$$

where, A_{ij} is the (i, j) element of the loading matrix A and the parameters $\left(\sqrt{\frac{\lambda}{2}}, \sqrt{2\gamma} > 0\right)$ are the scale parameters of the Laplacian distribution. Since the Laplacian distribution is not a conjugate to the Gaussian distribution, defining the priors for the

loadings $A_{i,j}$ and noise ε_i as presented by the equations above may lead to an intractable formulation. Fortunately, the Laplacian distribution can be represented as an infinite superposition of Gaussian distributions by introducing the intermediate variables $\Lambda_{i,j}$ and Γ in the following manner:

$$p(A_{i,j}|\lambda) = \int p(A_{i,j}|\Lambda_{i,j}^{-1})p(\Lambda_{i,j}|\lambda)d\Lambda_{i,j}$$
$$= \int \sqrt{\frac{\Lambda_{i,j}}{2\pi}} exp\left\{-\frac{\Lambda_{i,j}A_{i,j}^2}{2}\right\} \frac{1}{\lambda\Lambda_{i,j}^2} \exp\left\{-\frac{1}{\lambda\Lambda_{i,j}}\right\} d\Lambda_{i,j} \quad , \quad i \in \{1, \dots, p\}, j \in \{1, \dots, q\} \quad (3.4)$$

$$p(\varepsilon_i|\gamma^{-1}) = \int \sqrt{\frac{\Gamma\gamma}{\pi}} \exp\{-\Gamma\gamma\varepsilon_i^2\} p(\Gamma)d\Gamma \quad , \quad i \in \{1, \dots, p\}. \quad (3.5)$$
$$\text{and } p(\Gamma) = \frac{1}{2\Gamma^2} \exp\left\{-\frac{1}{2\Gamma}\right\}$$

In equations (3.4) and (3.5) we decompose the Laplacian distribution into a two-level hierarchy. The first level is to impose Gaussian distribution priors on $A_{i,j}|\Lambda_{i,j}^{-1}$ and $\varepsilon_i|\Gamma^{-1}$. The second level is to impose inverse Gamma distributed hyper-priors on the intermediate variables $\Lambda_{i,j}$ and Γ with their respective scale parameters $\frac{2}{\lambda}$ and $\frac{1}{2\gamma} > 0$, while the shape parameter is set to 1. To visualize how the variables in this reformulated model relate, Figure 3.2 shows a complete graphical representation. It is worth noting that if we marginalize out the first level priors $p(\Lambda_{i,j})$ and $p(\Gamma)$ in equations (3.4) and (3.5) we retrieve equations (3.2) and (3.3), respectively. The prior for the latent variable z is left as

it is in the probabilistic PCA model as $z \in \mathbb{R}^q \sim \mathcal{N}(0, \mathbf{\Phi^{-1}})$, where $\mathbf{\Phi}$ is a diagonal covariance matrix. Finally we impose a $Gamma(a, b)$ prior on $\gamma = \sigma_\varepsilon^{-2}$.

Given the alternative priors presented in equations (3.2) and (3.3), the joint distribution of the observation set x, latent variable, loading matrix, hidden parameters and hyperparameters $\{\Theta \equiv z, A, \Lambda, \lambda, \Gamma, \gamma\}$ becomes:

$$p(x, \Theta) = \prod_{t=1}^{T} p(x_t - Az_t | A, \Theta) p(z_i) \times \prod_{i=1}^{p} p(\varepsilon_i | \Gamma^{-1}) p(\Gamma | \gamma) p(\gamma) \prod_{j=1}^{q} p\left(A_{i,j} \middle| \Lambda_{i,j}^{-1}\right) p\left(\Lambda_{i,j}\right) \tag{3.6}$$

We are interested in evaluating the posterior distributions of the hidden variables given the observations. However, the posterior distributions are computationally intractable because the marginal distribution $p(x)$ cannot be obtained analytically. To circumvent this issue, we utilize Bayesian Variational Inference as an approximation tool for estimating the posteriors. The next subsection gives an overview of this approximation procedure and how we implement it in our model.

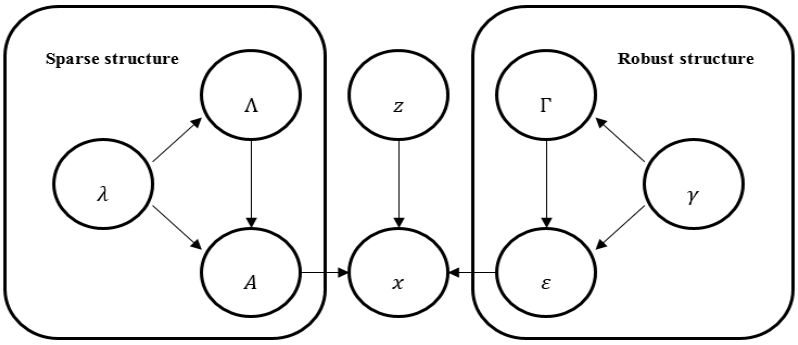


Figure 3.2 RSPCA graphical model (notation details in text). Arrows indicate conditional dependencies between model variables and parameters

3.3.2 Bayesian Variational Inference for RSPCA

The objective of Bayesian variational inference is to approximate the posterior distributions of the hidden variables $p(\Theta|x)$. Let $g(\Theta)$ represent this approximation, also referred to as the variational distribution (Wainwright and Jordan 2008). The KL-divergence between $g(\Theta)$ and $p(\Theta|x)$ is the difference between the log marginal likelihood $p(x)$ and a lower bound given by:

$$\mathcal{L}\big(g(\Theta)\big) = \int g(\Theta)\ln\frac{p(x,\Theta)}{g(\Theta)}d\Theta \leq \ln\int p(x,\Theta)d\Theta = \ln p(x) \tag{3.7}$$

Thus minimizing the KL divergence is equivalent to maximizing this lower bound, also known as the negative free energy in statistical physics. The objective then becomes to minimize the Kullback–Leibler (KL) divergence between $g(\Theta)$ and true posterior distribution $p(\Theta|x)$ as follows:

$$\max_{g(\Theta)} \mathcal{L}\big(g(\Theta)\big) \equiv \min KL(g(\Theta)||p(\Theta)\big) = \int g(\Theta)\ln\frac{g(\Theta)}{p(\Theta|x)}d\Theta \tag{3.8}$$

The next step is to postulate a simple parameterized family of distributions over $g(\Theta)$ such that it is tractable to evaluate the negative free energy while simultaneously obtaining a tight lower bound. Therefore, we assume that $g(\Theta)$ factorizes over all the parameters as follows:

$$g(\Theta) = g(z)g(A)g(\Lambda)g(\Gamma)g(\gamma) \tag{3.9}$$

The Bayesian variational inference procedure consists of two steps that are akin to the Expectation-Maximization (EM) algorithms. In the first step (E-step), the variational parameters are fixed at Θ, while the variational distribution g is updated to minimize the KL-divergence. The resulting variational posteriors as such:

$$g(\theta) \propto e^{E_{\Theta|\theta}[\ln p(x, \Theta|\theta)]} \tag{3.10}$$

Here $g(\theta)$ is the posterior of $\theta \in \Theta$ and $E_{\Theta|\theta}$ denotes the expectation with respect to all the parameters in the set Θ excluding the parameter θ, which is fixed while its respective posterior is calculated.

For the second step (M-step), we begin by fixing the updated variational posterior distribution obtained from the first step. Then we update the variational set of parameters Θ by maximizing the lower bound given the factorized variational distributions. The two steps are iterated, and the parameters are updated sequentially while the remaining are fixed. The next subsection summarizes the chosen family of variational distributions for the hidden parameters and hyperparameters, as well as their respected update equations.

3.3.3 Variational Posteriors and Update Equations

The best choice for the variational distribution of the latent variable $g(z_t)$ is the Gaussian density. The expectation over the variational posterior of a parameter θ will be represented by $\langle\theta\rangle$ for notation purposes. The variational posterior density will also be Gaussian with the following update equations for the mean $\langle z_t \rangle$ and variance Σ_z, respectively:

$$\langle z_t \rangle = \langle \gamma \rangle \langle \Gamma_t \rangle \Sigma_z \langle A \rangle^T x_t, \tag{3.11}$$

$$\Sigma_{z_t} = (\mathbf{\Phi} + \gamma \langle A^T \Gamma_t A \rangle)^{-1}, \tag{3.12}$$

Next, we look at the loading matrix A, where every row $A_{i.}$ is set to have independent Gaussian variational distribution. Hence the joint distribution of the rows of A is as follows:

$$g(A) = \prod_{i=1}^{p} g(A_{i.}), A_{i.} \in \mathbb{R}^q \sim \mathcal{N}(\langle A_{i.} \rangle, \Sigma_{A_{i.}}). \tag{3.13}$$

The respective update equations for the mean $\langle A_{i.} \rangle$ and variance $\Sigma_{A_{i.}}$ of the resulting Gaussian posterior are given by:

$$\langle A_{i.} \rangle = \langle \gamma \rangle \Sigma_{A_{i.}} \sum_{t=1}^{n} \langle z_t \rangle^T x_{.i} \tag{3.14}$$

$$\Sigma_{A_{i.}} = [\mathrm{diag}(\langle \Lambda_{i.} \rangle) + \gamma \sum_{t=1}^{n} \langle z_t^T \Gamma_t z_t \rangle]^{-1}. \tag{3.15}$$

As discussed in (Archambeau and Bach 2009, Gao 2008), the best choice of variational distribution for the hidden parameters Λ_{ij} is the generalized inverse Gaussian distribution given by:

$$g(\Lambda_{ij}) \sim GIG(\omega, \chi, \psi) \quad = \frac{\chi^{\omega} (\sqrt{\chi\psi})^{\omega}}{2K_{\omega}(\sqrt{\chi\psi})} {\Lambda_{ij}}^{\omega-1} exp\left\{-\frac{1}{2}\left(\chi {\Lambda_{ij}}^{-1} + \psi \Lambda_{ij}\right)\right\} \tag{3.16}$$

Here, $\omega = -\frac{1}{2}$ is the index, $\chi = l_{ij}$ and $\psi = \frac{2}{\lambda}$ are shape parameters, where l_{ij} is the j^{th} diagonal element of $\langle A_{i.}^T A_{i.} \rangle$. Moreover, $K_\omega(\cdot)$ represents the modified Bessel function of the second kind. The choice of the index and shape parameters given in equation (3.16) reflects the priors on $p(\Lambda_{ij})$ and $p(\mathrm{A}_{ij}|\lambda)$ described in subsection 3.3.1. Then, the update of Λ_{ij} is:

$$\langle \Lambda_{ij} \rangle = \sqrt{\frac{\chi}{\psi}} \frac{K_{\omega+1}(\sqrt{\chi\psi})}{K_\omega(\sqrt{\chi\psi})} = \sqrt{\frac{\lambda l_{ij}}{2}}. \tag{3.17}$$

Subsequently, we discuss the variational distribution of the hidden intermediate parameter of the noise Γ_t. The best choice for it is also the generalized inverse Gaussian distribution $GIG\left(-\frac{1}{2}, 1, \gamma\eta_t\right)$, as is the case for Λ_{ij}. This is because they both have the same two-level hierarchical structure, described in equations (3.8) and (3.9). Here, η_t is:

$$\eta_t = \frac{1}{p} \mathrm{tr}\left[(x_t - \langle \mathrm{A} z_t \rangle)(x_t - \langle \mathrm{A} z_t \rangle)^T + A \Sigma_{z_t} A^T\right], \tag{3.18}$$

where, $\mathrm{tr}(\cdot)$ refers to the trace of the enclosed matrix. Furthermore, the updates for Γ has the following form:

$$\langle \Gamma_t \rangle = \sqrt{\frac{1}{\gamma\eta_t}}. \tag{3.19}$$

Finally, the best variational distribution for the reciprocal of the error noise γ is still a $Gamma(a, b)$ with mean $\langle \gamma \rangle = \frac{a}{b}$. The update equations for the hyperparameters γ, λ and Φ respectively are:

$$a \longleftarrow a + \frac{np}{2}$$
$$b \longleftarrow b + \frac{1}{2}\sum_{t=1}^{n}\left[(x_t - \langle \mathrm{A}z_t\rangle)\langle\Gamma_t\rangle(x_t - \langle \mathrm{A}z_t\rangle)^T + \mathrm{tr}\left(\langle\Gamma_t\rangle A\Sigma_{z_t}A^T\right)\right], \tag{3.20}$$

$$\lambda \longleftarrow \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\Lambda_{ij}\,. \tag{3.21}$$

$$\Phi \longleftarrow \frac{1}{n}\sum_{i=1}^{n}\mathrm{diag}\left(\langle z_t\rangle\langle z_t\rangle^T + \Sigma_{z_t}\right). \tag{3.22}$$

3.3.4 RSPCA based Process Monitoring

The critical step in process monitoring is the evaluation of probability densities of all the variables. The probabilistic formulation of RSPCA that was presented in the previous subsections lays down the foundation for process monitoring. The probabilistic approach used in this chapter to model RSPCA assumes the densities of $p(z), p(\varepsilon)$ and $p(A)$ as discussed in subsection 3.3.1. What remains is to evaluate $p(x)$, as well as the posteriors $p(z|x)$ and $p(\varepsilon|x)$. For process monitoring using regular probabilistic PCA (Kim and Lee 2003), these densities are directly derived from using the probabilistic formulation of PCA given in equation (3.1). However, a similar direct approach is not possible for our proposed formulation as the posteriors cannot be obtained in a straightforward fashion. Therefore, we utilize the variational densities and their posteriors, which were discussed in subsection 3.3.3, to approximate the true densities.

3.3.4.1 Monitoring Latent Variables

We begin by discussing monitoring the latent variables z, which depending on the application may represent a fault pattern or a specific physical phenomenon that associates

multiple observation variables x. Our assumption for the distribution of the latent variables in the probabilistic model is that $z \sim \mathcal{N}(0, \mathbf{\Phi^{-1}})$. Since we cannot observe the latent variables directly, we propose to estimate them using their variational posterior densities $g(z)$. Given a new sample x_τ, the hypothesis test for whether the sample is in-control is as follows:

$$\begin{matrix} H_0: z_\tau = 0 \\ H_1: z_{\tau,i} \neq 0 \end{matrix}, \quad \text{For at least one } i \text{ such that } i \in \{1, \dots, q\} \tag{3.23}$$

Since we are dealing with Gaussian density for $z_{\tau,i}$, the test statistic is $\mathrm{X}_0 = \langle z_t \rangle^T \Sigma_z^{-1} \langle z_t \rangle$ and we propose to reject the null hypothesis H_0 given in (3.23) if $\mathrm{X}_0 = \left\| \langle \gamma \rangle \langle \Gamma_t \rangle \Sigma_z^{1/2} \langle A \rangle^T x_t \right\|^2 > \chi^2_{(1-\alpha,q)}$.

3.3.4.2 Monitoring the noise variable

Monitoring the latent variable is useful for detecting out of control instances, which is the case only when the instance is in accordance with the RSPCA model developed in section 3.3.1. This is where monitoring the noise variable comes into play. It identifies instances that do not share the same subspace structure. This is similar to the Q-statistic that complements the hoteling-T^2 statistic. In a similar fashion to monitoring the latent variable, the distance between the incoming sample and the model will be estimated using the mean of the variational posterior density and the hypothesis test then becomes:

$$\begin{matrix} H_0: \varepsilon_\tau = 0 \\ H_1: \varepsilon_{\tau,i} \neq 0 \end{matrix}, \quad \text{For at least one } i \text{ such that } i \in \{1, \dots, q\}. \tag{3.24}$$

The test statistic is then $\mathrm{X}_0 = \gamma\langle\varepsilon_\tau\rangle^T\langle\varepsilon_\tau\rangle$ and we propose to reject the null hypothesis given by (3.24) if $\mathrm{X}_0 = \left\|\langle\gamma\rangle^{-1/2}\langle\Gamma_t\rangle^{-1/2}(x_t - \langle \mathrm{A}z_t\rangle)\right\|^2 > \chi^2_{(1-\alpha,p)}$.

3.3.5 Fault Diagnosis

Once an incoming sample has been identified to be an out-of-control instance, it is desired to isolate the responsible variables for this irregularity. This diagnostic step can be quite difficult since our detection of the out-of-control instance is based on the hidden latent variables rather than the observed variables themselves. Therefore, the diagnosis method should be able to distinguish the observable variables that contribute towards the irregularity detected in the latent variable subspace.

The diagnostic procedure can be decomposed into the following steps: (1) identifying the out-of-control latent variable $\left(z_j, j \in \{1, \dots, q\}\right)$ and (2) determining the set $S \in \mathbb{R}^s$ of observable variables that contribute to the identified latent variable z_j. It is important to mention that for the remainder of this study we assume that only one latent variable can go out-of-control at any given time in the steps mentioned previously. This could be interpreted as each fault type being associated with a single latent direction.

Isolating the out of control latent variable z_j can be achieved via decomposition methods for the Hotelling T^2 statistic of the hypothesis test (3.23). The most common method would be the Mason-Tracey-Young (MTY) method that was proposed in (Mason

et al. 1995). This decomposition relaxes to finding the latent variable z_j that has a significant deviation from the mean. In other words, z_j such that $(\langle\Gamma\rangle\Sigma_z\langle A\rangle^T x_\tau)_j^{\,2} > \frac{\tau+1}{\tau}F_{1,\tau-1}$.

Determining the set of observable variables that contribute to the out-of-control latent variable z_j is not as straightforward. This is because A_{ij} is a random variable, and finding the set of $x_i \in S$ such that $A_{ij} \neq 0$ corresponds to the following hypothesis test:

$$\begin{aligned} H_0&: A_{ij} = 0 \\ H_1&: A_{ij} \neq 0 \end{aligned}. \tag{3.25}$$

The hypothesis in (3.25) is rejected when the test statistic $\left[\left(\Sigma_{A_{i.}}\right)_{jj}\right]^{-1/2}\left[\langle\gamma\rangle\Sigma_{A_{i.}}\sum_{t=1}^{n}\langle z_t\rangle^T x_{.i}\right]_j \leq t_{\tau-1}$. Here, $\left(\Sigma_{A_{i.}}\right)_{jj}$ is the jth diagonal element of $\Sigma_{A_{i.}}$, which corresponds to the marginal variance of A_{ij}. For all A_{ij} $(i = 1, \dots, p)$ such that the hypothesis is rejected, we conclude that the observed variable $x_i \in S$ contributes to the detected out-of-control z_j instance.

3.4 Simulations

This section presents the results of simulation studies to validate our proposed method and to test its monitoring performance. The first subsection 3.4.1 discusses the method in which we generate the simulation data. The following subsection 3.4.2 illustrates the ability of the proposed methodology to accurately recover the loading matrix and compares it to the state-of-the-art methods, which also serves as a verification step. The final subsection

3.4.3 evaluates the monitoring capability using the proposed methodology, while comparing it to other benchmark dimension reduction techniques.

The robust and sparse properties of our proposed methodology are evaluated in this section against state of the art techniques from the literature. Two benchmark methods are considered in the simulated experiments. The first method by Croux et al. (2013) will be denoted as SRPCA, and the second being ROSPCA (Hubert et al. 2016). Initially, we describe the simulated data generation procedure in subsection 3.4.1.

3.4.1 Data Generation

The generative model given by (3.1) is the base of the data generation method used in the following simulations. We adopt a setup consistent with the one described in (Hubert et al. 2016) for generating the data. First the loading matrix A is generated. To promote sparsity in the covariance of the observation variables x, the columns of A are designed to be sparse in a block-wise fashion with K blocks. Block B_k has cardinality $|B_k| = b_k$ such that the corresponding loading $A_{i,j}$ for all x_i and z_j becomes:

$$A_{i,j} = \begin{cases} -\frac{1}{\sqrt{b_k}} & \text{if } x_i \text{ and } z_j \in B_k \text{ for some } k \\ 0, & \text{otherwise} \end{cases}. \tag{3.26}$$

The nodes represent the observable and latent variables, while edges represent a non-zero element of the loading matrix. The cardinality of the first two blocks are chosen to be the same (i.e. $b_1 = b_2 = b$), while the remaining blocks are unit blocks ($b_k = 1$ for all $k > 2$). The total number of blocks, which is also the number of latent variables, is thus $K = p - 2b + 2$.

Figure 3.3 provides a network visualization of the blocks. Next, the latent variables z_j are generated from a normal distribution with zero mean and a diagonal covariance matrix Σ_z. The variances of z_1 and z_2 are set to be significantly larger than the remaining latent variables so the principal components corresponding to them can be identified as the first and second, respectively. Finally, white noise is added to normal observation and $100\delta\%$ of the observations are replaced by outlier observations. Outliers are independently generated from a multivariate normal distribution with mean μ_{out} and diagonal covariance matrix $\sigma_{out}^2 \boldsymbol{I}_p$. The outliers are generated such that they do not follow the correlation structure of the normal observations which will emphasize the need for robustness.

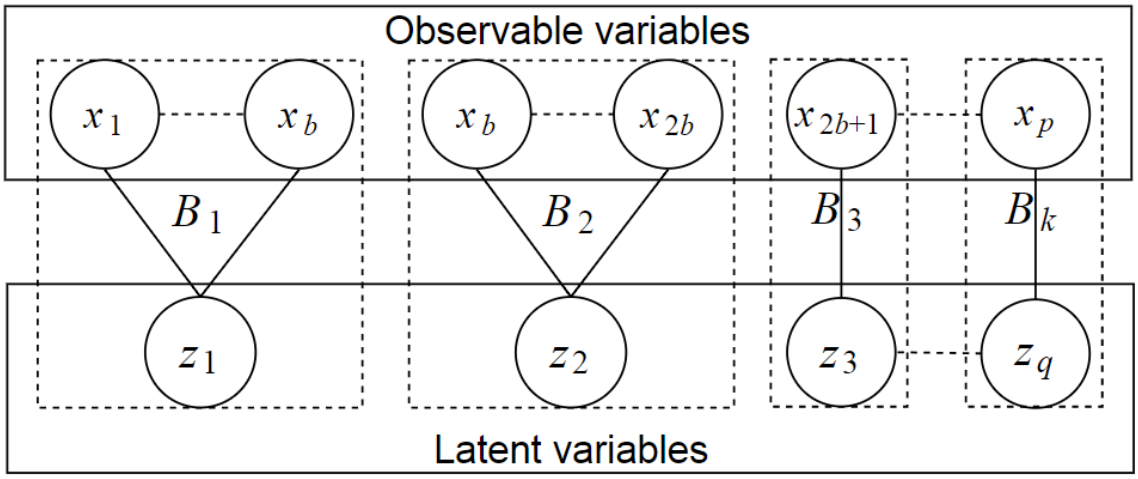


Figure 3.3 Network visualization of the simulation data generation blocks

3.4.2 Loading Matrix Recovery Experiment

In this subsection, we evaluate the recovery of the loading matrix by generating a set of data as described in the previous subsection. We consider a high-dimensional setting with $p = 500$ and $n = 50, 100, 500$. Only cases $n \leq p$ where considered as the difference

in the accuracy of retrieving the principal components via the different methods becomes negligible. Moreover, having the sample size be smaller than the dimension of the data is the challenging scenario of interest in this study. This is consistent with benchmark studies and it is also similar to the motivating Buckypaper manufacturing process monitoring problem, where the dimension of Raman spectra is 512. The first two blocks have size $b = 20$ and the variances of the latent variables $\mathrm{diag}(\Sigma_z) = [233, 49, 4(422 \text{ times}), 2(19 \text{ times}), 0.4(19 \text{ times})]$. Outliers with mean $\mu_{out} = 25(0, -4, 4, 2, 0, 4, -4, 2, \ldots \text{(for the first two blocks)}, 3, -3, \ldots, 3, -3)^T$, variance $\sigma_{out}^2 = 20$ and proportions $\delta = 0.1, 0.2, 0.3$.

We simulate 100 datasets for each experimental setting to keep computations practical. To compare the robustness of our proposed RSPCA method against SRPCA and ROSPCA, we utilize the average deviation angle as was used in the benchmark studies. This computes a measure between the estimated subspace and the true subspace and results in an angle between 0 and $\frac{\pi}{2}$, which is then normalized to produce a measure between 0 and 1. Values closer to 0 are desired as they represent a closer estimate. The tuning parameter that controls sparsity for SRPCA and ROSPCA is chosen based on the BIC criterion proposed in the respective study. For ROSPCA, the parameter determining the degree of robustness, which constitutes a lower bound on the number of normal observations, is set to 0.5 for maximal robustness as suggested in (Hubert et al. 2016). The average deviation angle (standard deviations) results from the experiments are summarized in Table 3.1.

Table 3.1. Average deviation angle (standard deviation) values of extracted PCs

$\boldsymbol{\delta}$	0.1			0.2			0.3		
$\boldsymbol{n}$	50	100	500	50	100	500	50	100	500
RSPCA	0.61 (0.09)	0.36 (0.06)	0.15 (0.03)	*0.63* (0.12)	*0.40* (0.09)	*0.16* (0.05)	*0.69* (0.14)	*0.45* (0.15)	0.18 (0.08)
ROSPCA	*0.59* (0.11)	*0.34* (0.08)	*0.14* (0.04)	0.64 (0.14)	0.42 (0.10)	0.17 (0.05)	0.70 (0.17)	0.46 (0.14)	*0.18* (0.07)
SRPCA	0.71 (0.18)	0.44 (0.13)	0.15 (0.06)	0.75 (0.21)	0.51 (0.15)	0.18 (0.08)	0.82 (0.32)	0.58 (0.13)	0.19 (0.17)

The simulation results indicate that the estimation of the principal components improves as n increases in terms of both bias and variance. Our proposed probabilistic approach appears to yield better results than SRPCA, while being competitive with ROSPCA and even slightly outperforming it in the case of moderate outliers ($\delta = 0.2$).

We use the *zero measure* to compare the three techniques in correctly identifying the sparse structure. The *total zero measure* is the proportion of loadings correctly identified as 0 or nonzero. For SRPCA and ROSPCA, an element of a loading matrix is considered to be 0 if its absolute value is smaller than 10^{-5}. While for our proposed RSPCA method, we test the hypothesis in (3.25) at significance level 0.01 to determine whether an element is 0 or not.

Figure 3.4 illustrates the total zero measure for RSPCA against the benchmark methods. It can be seen that our proposed probabilistic approach is superior in distinguishing the sparse structure of the principal subspace because the probabilistic model allows for a better way to discern zero loadings via hypothesis testing.

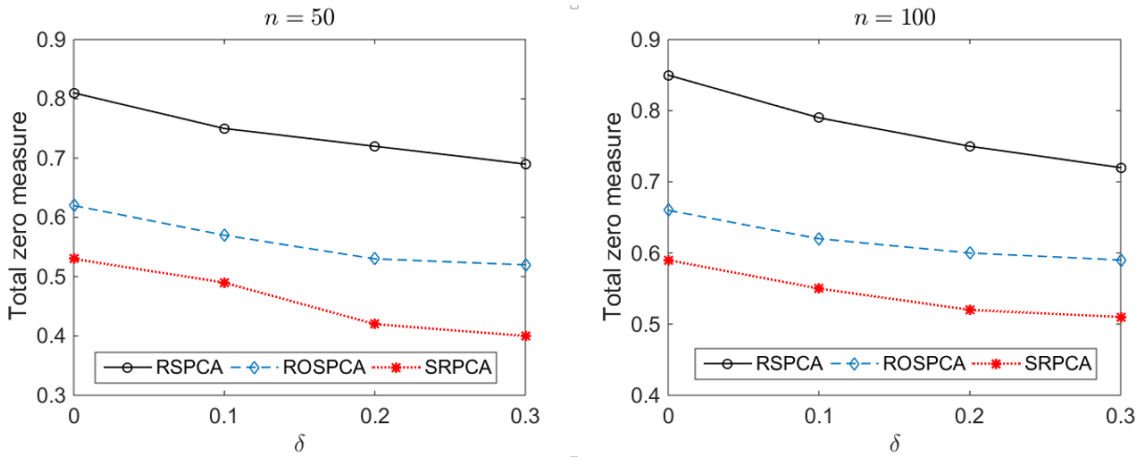


Figure 3.4 Total zero measure of RSPCA vs. benchmark methods ROSPCA and SRPCA

3.4.3 Monitoring Performance

This subsection is aimed towards testing the performance of our method in detecting changes that occur while monitoring a process. To better demonstrate the need for a monitoring procedure with both robust and sparse properties, we compare the detection delay to that of classical PCA, sparse PCA and robust PCA. The implementation of the other variations is based on modifications to our own method to remove robustness, sparsity, or both during the extraction of principal components. This self-implementation is similar to the proposed approaches in the literature (Zeng et al. 2017, Ge and Song 2011, Kim and Lee 2003). In control data sets are generated using the same method from the previous subsections, while out-of-control observations are generated by shifting the first latent variable z_1 by a range from 0 to 2 standard deviations from the mean. The principal components are learned from $n = 500$ in control observations with outliers with contamination proportion $\delta = 0.1, 0.2$. The average detection delays from 100 iterations are summarized in Figure 5, where the in-control average run length is set to 200.

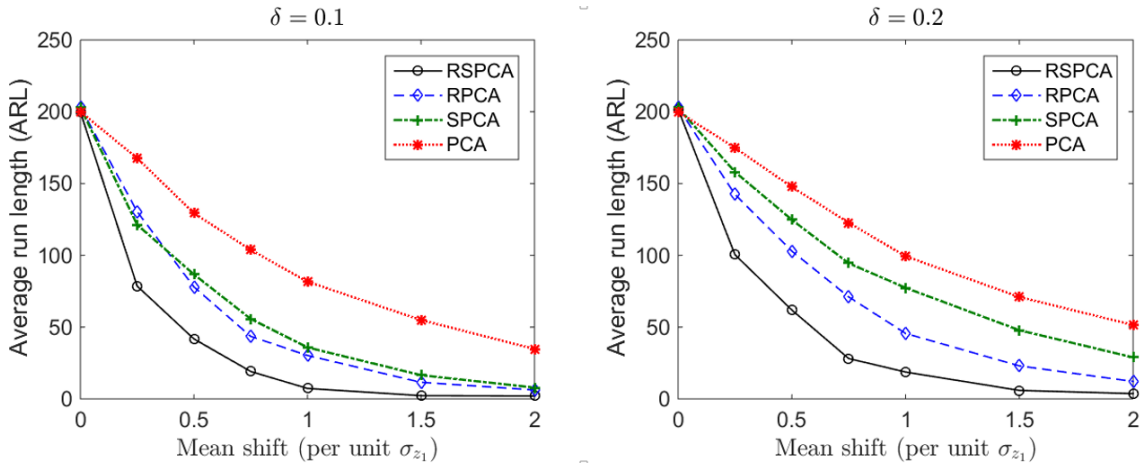


Figure 3.5 Illustration of the detection delay

From the results of Figure 3.5, it can be seen how the detection delay has been decreased significantly when using our proposed approach, which considers both robustness and sparsity to improve the change detection monitoring capability. At the lower contamination level $\delta = 0.1$ the performance of RPCA and SPCA is relatively similar especially at very small mean shifts. However, the advantages of robust estimation becomes more clear at the higher contamination level $\delta = 0.2$, which is to be expected. Classical PCA performs relatively poorly in all settings since it does not take into account the sparse structure or the outliers in the training data.

3.5 Case Study

In this subsection we test our proposed methodology in addressing the challenges of monitoring the production process of continuous CNTs buckypaper using inline Raman spectroscopy. We aim to show that the sparsity and robust properties of our proposed method can address the sparse peak locations and the complex noise structure.

The data we use in our case study is from a surrogated Raman spectra from a practical experiment. More details on the experiment and the data acquisition can be found in (Yue et al. 2018). Figure 3.6 illustrates the Raman spectra obtained from the experiments. The highlighted regions correspond to the sparse segments where defects occur. The first highlighted segment represents the D-band while the third is the G-band. These bands contain relevant quality information such as molecular defects in the CNT structure as well as functionalization (Cheng et al. 2010). Therefore, it is important to be able to detect irregularities in these segments when monitoring the whole profile. Defects, which are not associated with either the D-band or G-band, can also occur in other regions such as the middle region highlighted in the figure. From the zoomed in box of profiles, we can note the existence of outlier observations (solid blue profiles) with excessive noise that appear to mask the defects.

We begin the monitoring procedure by extracting the principal components. The components that most explain the variance in each of the sparse segments are respectively shown in Figure 3.7 for the different PCA methods. We can see that our proposed robust and sparse procedure successfully isolates the sparse segments similar to sparse PCA. While the ones obtained by regular PCA and robust PCA are mixed with other regions. Moreover, the inherent robustness property of our method provides a better representation of the segments in their respective principal component without dilution from the other segments. This allows our method to better isolate the segments when compared to sparse PCA.

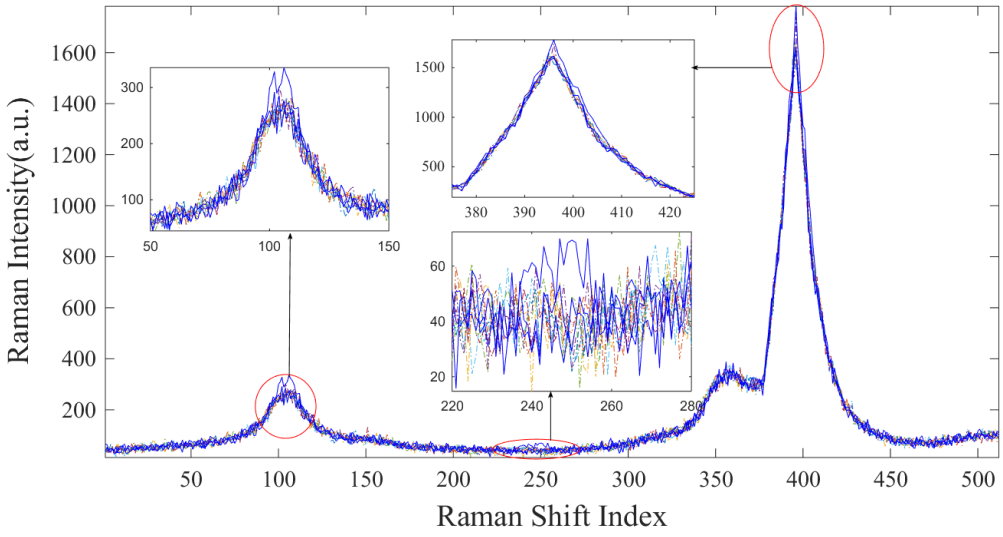


Figure 3.6 Illustration of the Raman spectra data

Next, we project the original data on the extracted principal components and test whether the profiles are in control or out of control. Figure 3.8 shows the plots for the projection of a sample of the original data onto the extracted components (PC scores) for a representative sample. While Figure 3.9 demonstrates the mean shift in the sparse segments of the out-of-control profiles. Table 3.2 summarizes the fault detection delay results from 1000 iterations.

Table 3.2 Detection delay comparison between RSPCA and the other PCA techniques

	Defect 1	**Defect 2**	**Defect 3**
Proposed RSPCA	*4.1(3.4)*	*3.2(3.4)*	*2.5(1.6)*
PCA	25.6(10.1)	17.2(6.2)	19.7(5.6)
Sparse PCA	20.1(5.8)	18.9(4.3)	10.6(3.9)
Robust PCA	12.1(6.5)	16(5.1)	9.6(4.2)

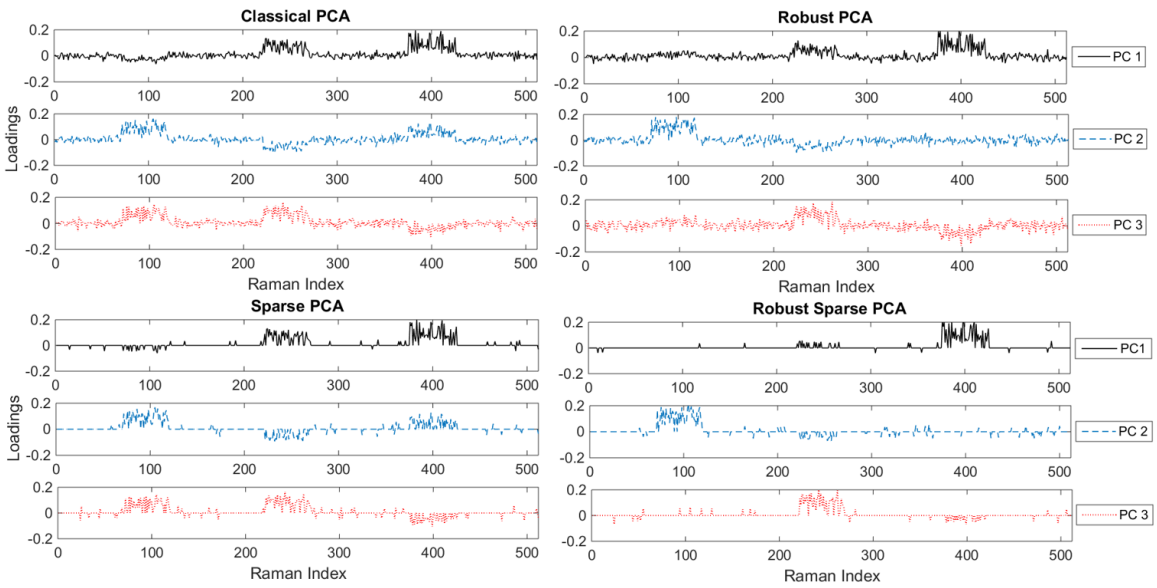


Figure 3.7 Demonstration of extracted significant principal components of the Raman data

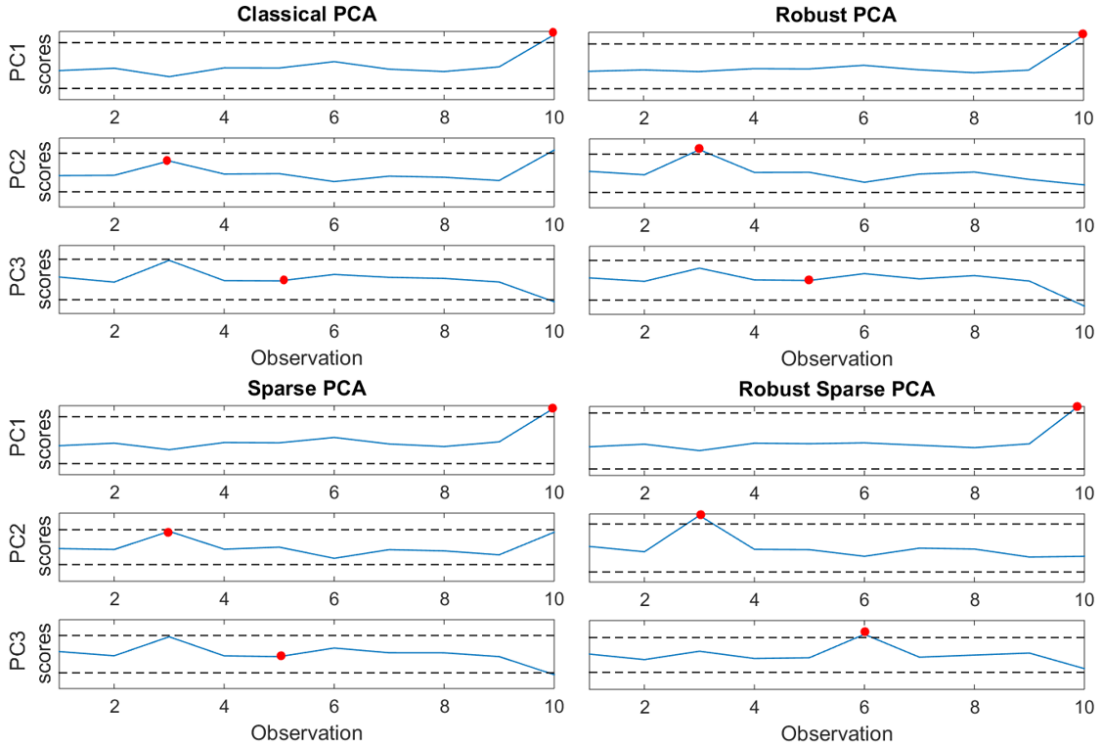


Figure 3.8 Projection of representative data on the PCs from RSPCA and other benchmarks

Note that the defective profiles can be clearly spotted by the projection to the principal component corresponding to the defective segment using RSPCA. These projections are marked by the red dots in Figure 3.8. The red projections are highly pronounced by our proposed method when compared to the remaining PCA techniques. This is reinforced from the results of Table 3.2, where the detection delay of the robust sparse PCA is significantly better than its counterparts for defects occurring in any of the regions highlighted in Figure 3.9.

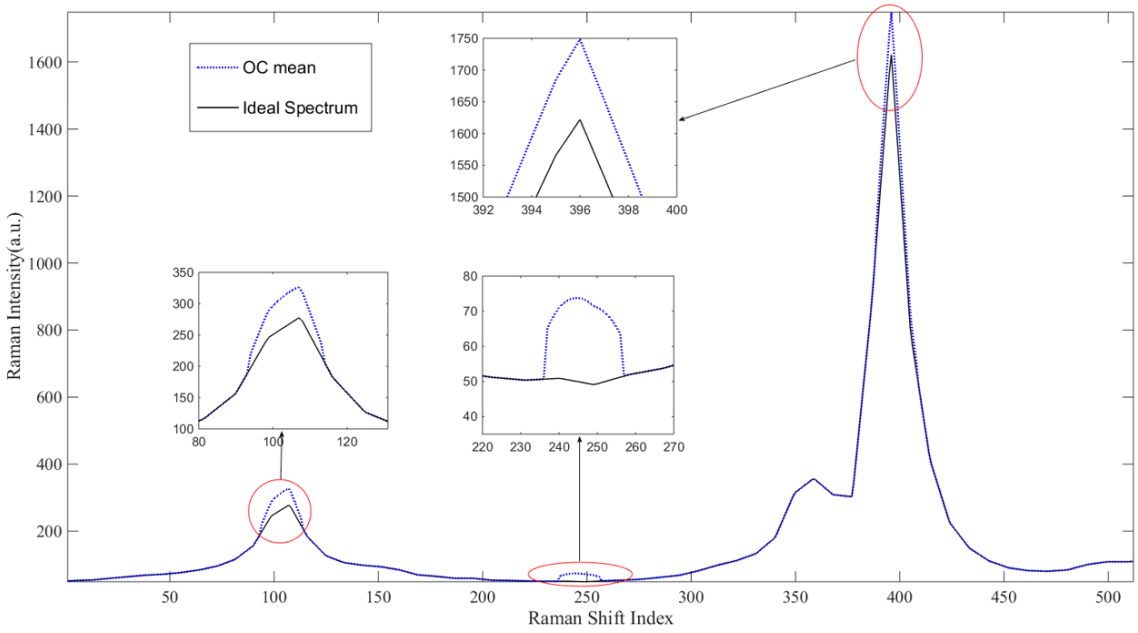


Figure 3.9 Illustration of the out-of-control shift in the sparse segments of the profiles

3.6 Conclusion

Change detection in processes that produce high dimensional data streams has become crucial with the advancement in sensing technologies. This chapter proposes a novel method that exploits the spatial structure of the data streams while simultaneously reducing the dimension. This is achieved by using a probabilistic model to extract robust

sparse principal components. This is advantageous in the sense that it allows for consistent and flexible modelling of the data streams based on the underlying spatial structure even in the event of noise and outliers.

This research introduces a new way to deal with high dimensional data streams. It can be used as a feature selection methodology that isolates the streams and capture the general structure, such that they can be monitored without being affected by the noise of insignificant streams or outliers. Moreover, the probabilistic approach in modelling provides a direct medium for change detection and diagnosis.

We tested the performance of our RSPCA method against other benchmark robust and sparse PCA based techniques. The results from both the simulation and the case study demonstrate the effectiveness of our proposed procedure in dealing with data with sparse irregularities and outliers. The case study from Raman spectroscopy of buckypaper manufacturing highlights the capability of the method to isolate sparse segments from high dimensional profiles while simultaneously minimizing the effect of outlier noise.

Our proposed approach mainly relies on robustly obtaining the sparse principal components based on the structure of the observation variables, and as is with regular PCA, these components are linear combinations of the original data. Kernel PCA (Schölkopf et al. 1997) can accommodate non-linear structures that may be embedded into the original data. Extensions to nonlinear relations was not discussed in this chapter but is interesting to explore carefully in its own right in future research.

CHAPTER 4. ADAPTIVE ROBUST SPARSE PROBABILISTIC PRINCIPAL COMPONENT ANALYSIS FOR PROCESS MONITORING

4.1 Introduction

Many modern manufacturing processes have different operating conditions that can affect the monitoring efficacy when not accounted for. This could be a result of varying process settings, materials and equipment. Modeling these processes requires a dynamic description of the process in real time. This reduces the false alarm rate that would otherwise result from a static representation of the operating conditions. One challenging aspect of having an adaptive modelling technique is to handle the trade-off between adjusting the model based on novel observations while avoiding overfitting issues, especially in the event of outliers.

In Chapter 3, we presented robust sparse probabilistic PCA modelling procedure to address the monitoring of the manufacturing process of carbon nanotubes (CNTs) buckypaper in real time using in-line Raman spectroscopy. The modelling and monitoring scheme was shown to be capable for the inline monitoring of Raman spectrums when the process settings are time invariant. However, such a static model cannot handle the changing operating conditions of the manufacturing process, which can result in different noise and correlation structures, as well as different signal intensities affecting the peak magnitudes as shown in Figure 4.1.

There are several sources for variation in Raman spectrums. One source of variation is the acquisition frequency of signals (Yue et al. 2017b). Characterization of an inline Raman spectrum takes multiple scans of at least 10 seconds per scan. The quality of the acquired spectrum, in terms of signal-to-noise ration (S/N) depends on the duration of the scans. Shorter data acquisition times can result in lower S/N ratios due the rapidly moving samples. Furthermore, the shorter they are, the less resulting material heterogeneity dependent noise. During a continuous test, material heterogeneity augments the signal differences among locations that are further apart. Therefore, a transition in signal properties is to be expected as the manufacturing process continues. Finally, externally generated noise from external light sources, such as fluorescent room lighting, may add to the variation of the acquired spectrums. Figure 4.1 illustrates acquired Raman spectrums from two operating periods (red, blue), where both the acquisition frequency and signal intensity are higher in the second operation settings.

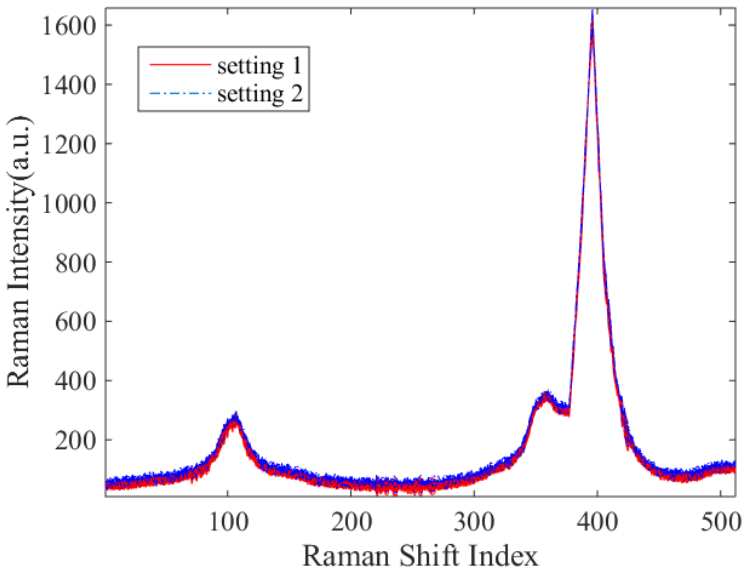


Figure 4.1 Illustration of the Raman spectra data

To address these issues, adaptive modelling and monitoring techniques need to be implemented. Preceding work on the problem of obtaining adaptive principal components as well as sparse or robust variations include (Liu et al. 2015a, Li et al. 2003, Li et al. 2000), which will be used as benchmarks in the simulation experiments and case study. While the aforementioned work may handle the nonstationary challenges of the problem of interest, a both robust and sparse adaptive procedure is still required.

This chapter introduces a procedure to dynamically model the process by developing an adaptive implementation of the robust sparse principal component analysis proposed in Chapter 3. The method sequentially solves the variational inference problem of the probabilistic model using stochastic optimization. The novelty of the proposed approach is in the adaptive estimation of the principal components using the information from new observations. Additionally, an adaptive learning rate that increases as the novelty in the data increases, while still being robust to outliers, is proposed. This allows the monitoring procedure to adjust to new process behavior that would otherwise trigger false alarms.

The rest of this chapter is organized as follows: In Section 3.2, we provide an overview of relevant topics in the literature followed by an introduction to stochastic variational inference. This will offer sufficient background information for the methodology developed in section 3.3. Sections 3.4 and 3.5 demonstrate the effectiveness of our proposed procedure through simulations as well as a case study on the monitoring of inline Raman spectroscopy for buckypaper manufacturing. Finally, the chapter is concluded with a discussion of the major findings of our proposed modelling and monitoring scheme in section 3.6.

4.2 Literature Review

This section is split into three main subsections. The first subsection 4.2.1 presents a brief overview of some of the available adaptive variations of principal component analysis in the literature. The second subsection 4.2.2 provides a review of the application of stochastic variational inference, which is a stochastic optimization technique used for dynamically solving probabilistic models. This will facilitate the discussion of our proposed methodology in section 4.3.

4.2.1 Adaptive Principal Component Analysis

Principal Component Analysis (PCA) is a commonly used dimension reduction technique that projects the original observed data onto a low dimensional subspace (Jolliffe 2011). While it has been used effectively for process monitoring and fault detection, PCA has several drawbacks that limit its implementation to many advanced manufacturing processes. One major disadvantage of PCA is the fact that it is an offline modelling method that is invariant and therefore cannot be used for time varying systems.

To address this issue, several dynamic variations of PCA have been proposed in the literature. Dynamic and adaptive variants of PCA differ in the specific issue that they try to address. Two main issues that have been explored in the literature are auto-correlated and nonstationary systems. A good overview of methods addressing these limitations can be found in (Rato et al. 2016).

More relevant to our interest in modelling nonstationary systems, recursive principal component analysis (Li et al. 2000) updates the principal components by

incorporating new observations to update the mean and covariance matrix. In addition, moving window principal component analysis (Wang et al. 2005) uses a moving window of fixed size to update the principal components. Several methods have been developed to improve the efficiency of the updates including (Portnoy et al. 2016, Jeng 2010). To handle missing data, a recursive method based on probabilistic principal component analysis was proposed by Zhang et al. (2015). In order to improve the interpretability, an adaptive sparse principal component analysis method was developed by Liu et al. (2015b). Additionally, methods that attempt to recursively estimate robust principal components include (Lois and Vaswani 2015, Qiu et al. 2014, Feng et al. 2013, Li et al. 2003).

The main limitation of the aforementioned methods is that they are not able to address both sparsity and robustness simultaneously within their recursive formulations. This chapter proposes a method that addresses this issue by developing a recursive solution to a probabilistic robust sparse principal component analysis model. The recursive procedure takes advantage of stochastic optimization in the variational inference framework, which is introduced in the following subsections.

4.2.2 Stochastic Variational Inference (SVI)

Stochastic variational inference is a method developed to optimize the ELBO by following noisy estimates of the gradient (Hoffman et al. 2013). Traditional variational inference as described in the previous chapter, solves the ELBO via coordinate ascent by iterating between computing the optimal local parameters given the current setting of the global parameters and then updating the global parameters given the computed local parameters. At each setting of the global parameters, the ELBO is guaranteed to increase

and the method iterates until it converges to a local optimum. This iterative procedure can be inefficient when the training data is very large as the local parameters have to be computed for each data point. Stochastic variational inference allows for iteratively updating the approximate posteriors by repeatedly subsampling a data point and finding a noisy estimate for the posteriors.

This method is dynamic in the sense that the model is updated with every sampled data point. However, it was mainly developed for offline modelling, where the sample pool is very large such that regular variational inference is impractical. In subsection 4.3.1, we illustrate how this concept can be adopted into an online setting by considering novel observations as random samples for computing the noisy estimates. This shifts the dynamics of the method from an offline learning approach to an adaptive online learning approach.

4.3 Adaptive RS-PCA Methodology

Suppose that observed variables at any given time t are represented by the p dimensional vector $\boldsymbol{x_t} = (x_{1,t}, \dots, x_{p,t})'$. Our objective is to update an existing model representing the process at a past time, while simultaneously monitoring the data acquired to detect shifts in any component of the data streams. In our analysis we assume that the data can be projected on a lower dimensional subspace spanned by a set of latent variables $\boldsymbol{z_t} = (x_{1,t}, \dots, x_{q,t})'$. The objective of this chapter is to adaptively identify the lower dimensional subspace given an initial estimate of the subspace.

The next subsection 4.3.1 will demonstrate how this chapter uses stochastic optimization via variational inference to adaptively update robust sparse principal

components obtained from the probabilistic model introduced in Chapter 3. Subsection 4.3.2 discusses the implementation of a robust adaptive learning rate that allows for a higher learning rate from novel observations, while moderating the effect of outliers. Finally, subsection 4.3.3 describes how the proposed adaptive implementation of robust sparse principal component analysis can be utilized for online process monitoring.

4.3.1 SVI for Adaptive Robust Sparse PCA

In this subsection, we explain how stochastic variational inference can be utilized to adaptively approximate the variational posteriors in real time. The updates of the variational posteriors given in subsection 3.3.3 of the previous chapter represent the variational solution to optimizing the ELBO given a sample of n datapoints. Stochastic optimization (Robbins and Monro 1985) of the ELBO suggests that the update given a new observation $\boldsymbol{x_t}$, can be used as a noisy estimate of the gradient of the ELBO. This is because, each incoming observation is temporally independent and may be regarded as a subsampled point of a pool of data gathered from $t = 1,2, \dots, \infty$. Hence, stochastic optimization of the ELBO follows noisy estimates using each data acquisition time t with a learning rate (step-size) $\zeta_t \in (0,1)$ through this sequence:

- Computing the variational posterior of the local hidden variables $\boldsymbol{z_t}$ given the new observation $\boldsymbol{x_t}$ according to the following equations:

$$\langle z_t \rangle = \langle \gamma_{t-1} \rangle \langle \Gamma_t \rangle \Sigma_z \langle A \rangle^T (x_t - m_{t-1}), \tag{4.1}$$

$$\Sigma_{z_t} = (\boldsymbol{\Phi} + \gamma_{t-1}\langle A^T \Gamma_t A\rangle)^{-1} \tag{4.2}$$

- Updating the intermediate local parameter Γ_t according to:

$$\langle \Gamma_t \rangle = \sqrt{\frac{1}{\gamma_{t-1}\eta_t}}. \tag{4.3}$$

- Computing an intermediate estimate for the global variational parameters $\widehat{\Theta}_t = \left(\hat{A}_t, \widehat{\Phi}_t^{-1}, \hat{\gamma}_t\right)$ by repeating a window (w) of observations $(\boldsymbol{x_{t-w+1}}, \boldsymbol{x_t})$ N times:

$$\begin{gathered}\langle \hat{A}_{i.} \rangle_t^T = \langle \gamma_{t-1} \rangle \Sigma_{A_{i.,t}} N \sum_{j=t-w+1}^{t} \langle z_j \rangle^T \left(x_{j,i} - m_{j-1,i}\right), \\ \Sigma_{A_{i.,t}} = \left[\Sigma_{A_{i.}}^{-1} t-1 + N \sum_{j=t-w+1}^{t} \langle z_j^T \Gamma_j z_j \rangle\right]^{-1}\end{gathered} \tag{4.4}$$

$$\widehat{\Phi}_t^{-1} = \text{diag}\left(\frac{1}{w}\sum_{j=t-w+1}^{t}\left(\langle z_j \rangle \langle z_j \rangle^T + \Sigma_{z_j}\right)\right), \tag{4.5}$$

$$\begin{gathered}\hat{b}_t = \hat{a} + \frac{1}{2}\sum_{j=t-w+1}^{t}\Big[\left(\boldsymbol{x_j} - m_{j-1}\right)^T \langle \Gamma_j \rangle \left(\boldsymbol{x_j} - m_{j-1}\right) - 2\langle \Gamma_j \rangle \left(\boldsymbol{x_j} - \right. \\ \left. m_{j-1}\right)\langle A_{i.,j} \rangle^T \langle z_j \rangle + \langle \Gamma_j \rangle \text{tr}\left(\langle z_j^T z_j \rangle \langle A_{i.,j}^T A_{i.,j} \rangle\right)\Big], \\ \hat{a}_t = \hat{a} + \frac{Nwp}{2}, \hat{\gamma}_t = \frac{\hat{a}_t}{\hat{b}_t}.\end{gathered} \tag{4.6}$$

- Updating the mean vector $\widehat{m}_t$ as follows:

$$\widehat{m}_t = \frac{1}{w}\sum_{j=t-w+1}^{t}\left(\boldsymbol{x_j} - \langle A\rangle^T\langle z_j\rangle\right), \tag{4.7}$$

- Updating the current estimate of the global parameters $\Theta_{t-1} = (A_{t-1}, \Phi_{t-1}^{-1}, m_{t-1})$ using the intermediate estimates $\widehat{\Theta}_t = \left(\hat{A}_t, \widehat{\Phi}_t^{-1}, \widehat{m}_t\right)$ with step size ζ_t as follows:

$$\Theta_t = (1-\zeta_t)\Theta_{t-1} + \zeta_t\widehat{\Theta}_t. \tag{4.8}$$

- Updating the global hyper-parameters $\theta_t = (\Lambda_t, \lambda_t)$ according to the following equations:

$$\langle\Lambda_{ij,t}\rangle = \sqrt{\frac{\chi}{\psi}}\frac{K_{\omega+1}(\sqrt{\chi\psi})}{K_{\omega}(\sqrt{\chi\psi})} = \sqrt{\frac{\lambda_{t-1}l_{ij,t}}{2}}, \tag{4.9}$$

$$\lambda_t = \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}\Lambda_{ij,t}\,, \tag{4.10}$$

where, $l_{ij,t}$ is the j^{th} diagonal element of $\langle A_{i.,t}^T A_{i.,t}\rangle$.

- Acquire the next data point $\boldsymbol{x_{t+1}}$ and repeat the sequence.

At each time t the method calculates an intermediate estimate of the global parameters by replicating the acquired data N times. Then, the update of the global parameters is computed as a weighted average of the current setting and the intermediate setting based

on the step size ζ_t. The convergence of the method depends on the choice of the step size. The conditions for convergence are as follows (Robbins and Monro 1985):

$$\sum_{t=1}^{\infty} \zeta_t = \infty, \quad \sum_{t=1}^{\infty} \zeta_t^2 < \infty. \tag{4.11}$$

Choosing the optimal sequence is an open research topic. A sequence with a fast decay may take longer to converge, while the opposite may result in volatile estimations. Moreover, we require a sequence with other properties to fit our modelling objectives. Since our system is nonstationary, the learning rate should adapt to the change in the system. We propose such a sequence in the next subsection and explore its properties.

4.3.2 Adaptive Robust Learning Rate

The problem of finding the optimal learning rate for estimating parameters in dynamic programming is a challenging problem that is highly dependent on the stochastic nature of the model. George and Powell (2006) provides a detailed review of methods for determining optimal deterministic and stochastic step-sizes.

Our problem is different from other stochastic problems in the literature in the sense that the transient period is unknown. Stochastic learning rates usually try to overcome initial transient behaviour which is followed by a steady state. In our ongoing monitoring procedure, the learning rates are expected to decay only when entering steady state and possibly increase when a transition is detected. This unpredictable behaviour makes it very challenging to guarantee an optimal learning rate that can also meet the sufficient conditions for convergence in equation (4.11).

The objective of this subsection is to develop a learning rate that is adaptive, robust and practical computationally. Suppose that the model that describes the current steady state of our process has the initial global parameters $\Theta_{(1)} = \left(A_{(1)}, \Phi_{(1)}^{-1}, m_{(1)}\right)$. These parameters could have been estimated from historical data. At an unknown time τ the process enters a transient state. In order to effectively monitor the process, the initial estimates of Θ need to be updated to reflect the transition to $\Theta_{(2)}$. To learn $\Theta_{(2)}$ we would require the observations $\boldsymbol{x_t}$ for $t = \tau, \tau + 1, \ldots, \infty$. Let $\Theta_{(2),\mathrm{T}}^{*}$ be the estimates for the global parameters obtained using observations $\boldsymbol{x_t}$ for time $t = \tau, \tau + 1, \ldots, \mathrm{T}$. Additionally, the stochastic updating method proposed in the previous subsection allows for sequential estimation $\widehat{\Theta}_{(2),t}$ of the global parameters. The learning rate should minimize the expected error Ξ between the stochastic update $\Theta_{(2),\mathrm{T}}$ and $\Theta_{(2),\mathrm{T}}^{*}$:

$$\Xi\left(\Theta_{(2),\mathrm{T}}^{*}, \Theta_{(2),\mathrm{T}}\right) = \left(\Theta_{(2),\mathrm{T}}^{*} - \Theta_{(2),\mathrm{T}}\right)^{T}\left(\Theta_{(2),\mathrm{T}}^{*} - \Theta_{(2),\mathrm{T}}\right). \tag{4.12}$$

A similar learning rate was proposed in (Ranganath et al. 2013), except there was only one process setting Θ. The adaptive learning rate that minimizes the square norm of the error can be shown to be (see APPENDIX B for derivation):

$$\zeta_{\mathrm{T}}^{*} = \frac{\Xi\left(\Theta_{(2),\mathrm{T}}^{*}, \Theta_{(2),\mathrm{T}-1}\right)}{\Xi\left(\Theta_{(2),\mathrm{T}}^{*}, \Theta_{(2),\mathrm{T}-1}\right) + \mathrm{tr}(\Sigma_{\Theta})}, \tag{4.13}$$

where, $\Sigma_{\Theta} = E\left[\left(\widehat{\Theta}_{(2),\mathrm{T}} - \Theta_{(2),\mathrm{T}}^{*}\right)\left(\widehat{\Theta}_{(2),\mathrm{T}} - \Theta_{(2),\mathrm{T}}^{*}\right)^{T}\right]$. This learning rate grows when the estimate $\Theta_{(2),\mathrm{T}}^{*}$ is far from the current sequential update $\Theta_{(2),\mathrm{T}}$ obtained using equation

(4.8). However, this learning rate does not account for outlier data that could inflate the learning rate and potentially delay convergence. To address this issue, we propose using the Huber's criterion as a substitute for $\Xi\left(\Theta^{*}_{(2),\mathrm{T}}, \Theta_{(2),\mathrm{T}-1}\right)$:

$$\mathcal{H}[\Xi(\xi)] = \begin{cases} \Xi(\xi) & \sqrt{\Xi(\xi)} \leq s\mathrm{M} \\ 2s\mathrm{M}\sqrt{\Xi(\xi)} - \mathrm{s}^2\mathrm{M}^2 & \sqrt{\Xi(\xi)} > s\mathrm{M} \end{cases}, \tag{4.14}$$

where, $s > 0$ is the scale parameter of the distribution. The parameter M dictates where the transition from quadratic to linear begins. The larger the value of M is the more similar the criterion is to the original least square loss. The choice of M controls the robustness of the criterion and $\mathrm{M} = 1.345$ was proposed to provide adequate robustness while being efficient for normally distributed data (Huber 1981). Furthermore, $\mathrm{tr}(\Sigma_{\Theta})$ can be used as an estimate for the scale parameter s^2. Hence, the robust adaptive learning rate becomes:

$$\zeta^{**}_{\mathrm{T}} = \frac{\mathcal{H}\left[\Xi\left(\Theta^{*}_{(2),\mathrm{T}}, \Theta_{(2),\mathrm{T}-1}\right)\right]}{\Xi\left(\Theta^{*}_{(2),\mathrm{T}}, \Theta_{(2),\mathrm{T}-1}\right) + \mathrm{tr}(\Sigma_{\Theta})}, \tag{4.15}$$

The issue with the proposed learning rate in equation (4.15) is that it requires the unknown quantities $\Theta^{*}_{(2),\mathrm{T}}$ and Σ_{Θ}. The quantity $\Theta^{*}_{(2),\mathrm{T}}$ can be represented as an expectation of the noisy gradient of the ELBO since $\mathrm{E}\left[\widehat{\Theta}_{(2),\mathrm{T}}\right] = \Theta^{*}_{(2),\mathrm{T}}$. Furthermore, the quantity $\mathrm{tr}(\Sigma_{\Theta})$ is the covariance of the noisy gradient $\mathrm{Cov}\left[\widehat{\Theta}_{(2),\mathrm{T}}\right]$. We propose to substitute these expectations with exponential moving averages, which is a solution first introduced in (Schaul et al. 2013) and has been adopted in similar dynamic programs by Ranganath et al. (2013).

Let η_{T} and H_{T} represent the moving averages used to approximate $\mathrm{E}\left[\widehat{\Theta}_{(2),\mathrm{T}} - \Theta_{(2),\mathrm{T}}\right]$ and $\mathrm{E}\left[\Xi\left(\widehat{\Theta}_{(2),\mathrm{T}}, \Theta_{(2),\mathrm{T}-1}\right)\right]$, respectively. Hence, using ϖ_{T} as the exponential window size we obtain the approximations:

$$\mathrm{E}\left[\widehat{\Theta}_{(2),\mathrm{T}} - \Theta_{(2),\mathrm{T}}\right] \approx \eta_{\mathrm{T}} = (1 - \varpi_{\mathrm{T}}^{-1})\eta_{\mathrm{T}-1} + \varpi_{\mathrm{T}}^{-1}\left(\widehat{\Theta}_{(2),\mathrm{T}} - \Theta_{(2),\mathrm{T}}\right) \tag{4.16}$$

$$\mathrm{E}\left[\Xi\left(\widehat{\Theta}_{(2),\mathrm{T}}, \Theta_{(2),\mathrm{T}-1}\right)\right] \approx \mathrm{H}_{\mathrm{T}} = (1 - \varpi_{\mathrm{T}}^{-1})\mathrm{H}_{\mathrm{T}-1} + \varpi_{\mathrm{T}}^{-1}\Xi\left(\widehat{\Theta}_{(2),\mathrm{T}}, \Theta_{(2),\mathrm{T}-1}\right) \tag{4.17}$$

Using equations (4.16) and (4.17) we can approximate ζ_{T}^{**} as follows:

$$\zeta_{\mathrm{T}}^{**} \approx \frac{\mathcal{H}[\eta_{\mathrm{T}}^{T}\eta_{\mathrm{T}}]}{\mathrm{H}_{\mathrm{T}}}, \tag{4.18}$$

where, the estimate of the scale parameter $s^2 = \mathrm{tr}(\Sigma_{\Theta})$ will be approximated by $(\mathrm{H}_{\mathrm{T}} - \eta_{\mathrm{T}}^{T}\eta_{\mathrm{T}})$. The moving averages can be initialized through Monte Carlo estimates of the expectations using the initial estimates $\left(\Theta_{(1)}, \theta_{(1)}\right)$ obtained from the historical data. To make the moving averages more reliable after large steps, the window size ϖ_{T} is updated as follows:

$$\varpi_{\mathrm{T}+1} = \max(\varpi_{\mathrm{T}}(1 - \zeta_{\mathrm{T}}^{**}) + 1, \varpi_1), \tag{4.19}$$

where ϖ_1 is the initial value of the window for the moving averages, and it is set as the size of the historical training data. The window is bounded from above to allow the method to reset after reaching steady state, otherwise the procedure will fail to detect future transitions after a long steady state.

Algorithm 4.1 provides an overview for the adaptive robust sparse principal component analysis proposed in the previous subsection, using the robust adaptive learning rate proposed in this subsection. The intermediate estimates of the global parameters $\widehat{\Theta}$ can be estimated using a batch of data instead of just one data point. Batch updating may result in better intermediate estimates, but finding the optimal batch size is out of the scope of this chapter and should be explored on its own.

Algorithm 4.1: Adaptive Probabilistic (RS-PCA)	
	Input: $\Theta_{(1)}, \theta_{(1)}, \varpi_0, \mathrm{H}_0, \eta_0, \zeta_0^{**}$
	For $t = 1, \ldots, \infty$
1	Obtain the observation vector $\boldsymbol{x_t}$
2	Compute estimates of local parameters with eq. (4.1) and (4.2)
2	Compute intermediate global parameters $\widehat{\Theta}_{\mathrm{t}}$ with eq. (4.4),(4.5) and (4.6)
3	Update the moving averages H_1 and η_1 with eq. (4.16) and (4.17)
4	Update the learning rate ζ_t with eq. (4.18)
5	Update the window size ϖ_t with eq. (4.19)
6	Update the global parameters Θ_t with eq. (4.8)
	End

4.3.3 Process Monitoring using Adaptive RSPCA

The monitoring and diagnostic procedure using the proposed adaptive implementation of the robust sparse principal component analysis follows suit to the one proposed in (Nabhan et al. 2019). However, the key difference is that at each data acquisition time, the decision making is performed on the latest update of the model parameters rather than a static model. Table 4.1 represents the hypothesis tests relevant to decision making during the monitoring and diagnostic stages.

Table 4.1 Hypothesis tests for the variables of the probabilistic model

	Hypothesis	Test statistic	Reject H_0
z_τ:	$H_0: z_\tau = 0$ $H_1: z_{\tau,i} \neq 0$, For at least one $i \in \{1, \dots, q\}$	$\mathrm{X}_0 = \left\| \langle \gamma \rangle \langle \Gamma_t \rangle \Sigma_z^{1/2} \langle A \rangle^T x_t \right\|^2$	$\mathrm{X}_0 > \chi^2_{(1-\alpha,q)}$
ε_τ:	$H_0: \varepsilon_\tau = 0$ $H_1: \varepsilon_{\tau,i} \neq 0$, For at least one $i \in \{1, \dots, p\}$	$\mathrm{X}_0 = \left\| \langle \gamma \rangle^{1/2} \langle \Gamma_t \rangle^{1/2} (x_t - \langle \mathrm{A} z_t \rangle) \right\|^2$	$\mathrm{X}_0 > \chi^2_{(1-\alpha,p)}$
A_{ij}:	$H_0: A_{ij} = 0$ $H_1: A_{ij} \neq 0$	$T_0 = \left[\left(\Sigma_{A_{i.}} \right)_{jj} \right]^{-1/2} \left[\langle \gamma \rangle \Sigma_{A_{i.}} \sum_{t=1}^{n} \langle z_t \rangle^T x_{.i} \right]_j$	$\lvert T_0 \rvert > t_{\tau-1}$

The step-by-step procedure for adaptive process monitoring using our proposed approach is given below. First, the monitoring procedure given an incoming observation $\boldsymbol{x_t}$ is as follows:

1) Check whether or not the estimate of the latent variables z_τ given the observation $\boldsymbol{x_t}$ is in control from the corresponding hypothesis test in Table 4.1.
2) Check if the observation is significantly different from the estimated model using the hypothesis test on the noise ε_τ from Table 4.1.
3) If neither one of the hypotheses in steps 1 and 2 were rejected then proceed to step 4, otherwise raise an alarm and move to the fault diagnostic steps.
4) Update the current estimates of the generative probabilistic model using algorithm 1 given the latest observation $\boldsymbol{x_t}$.
5) Acquire the next observation $\boldsymbol{x_{t+1}}$ and return to step 1

Second, given an out-control-signal from the previous monitoring steps, the diagnostic procedure is conducted as follows:

1) If the out-of-control signal was raised from monitoring the noise variables ε_τ, we conclude that the incoming data does not adhere to the current model. This requires revalidating the current model's adequacy in representing the data.
2) If the out-of-control signal was raised from monitoring latent variables z_τ, the Mason-Tracey-Young (MTY) method that was proposed in (Mason et al. 1995) is used to decompose the test and isolate the variables z_j that have significant deviations. If only one variable is considered at a time, this reduces to finding z_j such that $(\langle\gamma\rangle\langle\Gamma_\tau\rangle\Sigma_z\langle A\rangle^T x_\tau)_j^{\,2} > \frac{\tau+1}{\tau}F_{1,\tau-1}$.
3) Once the responsible latent variable z_j is determined, the contribution of observable variables x_i is tested using the hypothesis test for A_{ij} from Table 4.1.
4) The set of variables x_i, which the hypotheses of step 3 were rejected, are reported as contributing variables for the out-of-control signal.

4.4 Simulations

This section demonstrates the efficacy of the proposed adaptive robust sparse probabilistic PCA method proposed in this chapter. The objective of the simulation experiments is to assess the recovery of the new set of global parameters $\Theta_{(2)}$. The first subsection 4.4.1, describes the data generation procedure for the conducted experiments. The second subsection 4.4.2 provides a comparative study regarding the adaptive recovery of principal components using our proposed procedure against other benchmark methods from the literature.

4.4.1 Data Generation

The simulated experiments in this section are generated from the same probabilistic model of Chapter 3 given by equation (3.1). We adopt the experimental setup described by Nabhan et al. (2019) for generating the initial setting of the data. In the aforementioned setup, the loading matrix A is chosen to be block diagonal as follows:

$$A_{i,j} = \begin{cases} -\frac{1}{\sqrt{b_k}} & \text{if } x_i \text{ and } z_j \in B_k \text{ for some } k \\ 0, & \text{otherwise} \end{cases}.$$

Here, $A_{i,j}$ is the element corresponding to the loading that relates the observable variable x_i to the latent variable z_j. B_k denotes block k with cardinality b_k for $k = 1, \dots, K$. The dimension of the data is chosen to be $p = 500$ to be similar to the practical buckypaper monitoring case study, where the profiles consist of 512 variables. For both the initial and transitional settings, each block B_k contains one latent variable z_k. As for the observable variables in block B_k during the initial setting, $x_i \in B_1$ for $(i = 1, \dots, 20)$, $x_i \in B_2$ for $(i = 1, \dots, 20)$, and $x_{38+k} \in B_k$ for $(k = 3, \dots, 462)$. While during the transitional setting, $x_i \in B_1$ for $(i = 1, \dots, 10, 41, \dots, 50)$, $x_i \in B_2$ for $(i = 1, \dots, 20)$, $x_{8+k} \in B_k$ for $(k = 3, \dots, 12)$, and $x_{48+k} \in B_k$ for $(k = 13, \dots, 462)$.

The latent variables z_j are generated from a normal distribution with zero mean and a diagonal covariance matrix Σ_z such that: $\text{diag}(\Sigma_z) = [233, 49, 4(422 \text{ times}), 2(19 \text{ times}), 0.4(19 \text{ times})]$. Outlier instances contaminate $\delta \times 100\%$ $(\delta = 0.1, 0.2, 0.3)$ of the data and are generated form a normal distribution with shifted mean $\mu_{out} =$

$25(0, -4, 4, 2, 0, 4, -4, 2, \ldots, (\text{for the first two blocks}), 3, -3, \ldots, 3, -3)'$ and diagonal covariance matrix $\sigma_{out}^2 \boldsymbol{I}_p$, where $\sigma_{out}^2 = 20{,}40$.

4.4.2 Adaptive Loading Matrix Recovery Experiment

We simulate 100 datasets with sizes $n_{\text{initial}} = 500$ and $n_{\text{transitional}} = 50, 100, 500$. The initial datasets are used as a whole to estimate the initial model which is further updated with the transitional datasets via the different methods. Figures 2-4 illustrate the results from the experiments via box-plots that summarize the deviation angle between the estimated subspaces and the true subspace. This computes a measure of the angle between the estimated subspace and the true subspace resulting in a value between 0 and $\frac{\pi}{2}$, which is then divided by $\frac{\pi}{2}$ to give a value between 0 and 1. It is desirable for the estimated subspace to have a deviation angle close to 0, which represents an accurate estimate. Adjacent box-plots of similar color represent the results of the same method for $\sigma_{out}^2 = 20$ (left) and $\sigma_{out}^2 = 40$ (right).

We compare our proposed adaptive RSPCA method with robust (RPCA) and sparse (SPCA) variations of adaptive PCA (Liu et al. 2015b, Li et al. 2003). Additionally, we include the results of estimating the subspace using static RSPCA (Nabhan et al. 2019), which computes an estimate using only transitional data. This serves as a best case scenario that is generally unachievable in a practical case, where the transition point is not known. However, the inclusion provides insight as to how close the performance of the other methods are to an ideal case.

The figures show that our procedure is able to achieve close performance to the ideal estimation even when $n_{\text{transitional}}$ is small. Due to the presence of outliers, adaptive sparse PCA tends to have poor performance especially as δ increases. While the robust adaptive variation performs well, it still lags behind our method. This demonstrates the effectiveness of the sparse property in our procedure. Figure 4.4 illustrates the performance after obtaining a large transitional sample $n_{\text{transitional}} = 500$, where it can be seen that our method converges to the static procedure.

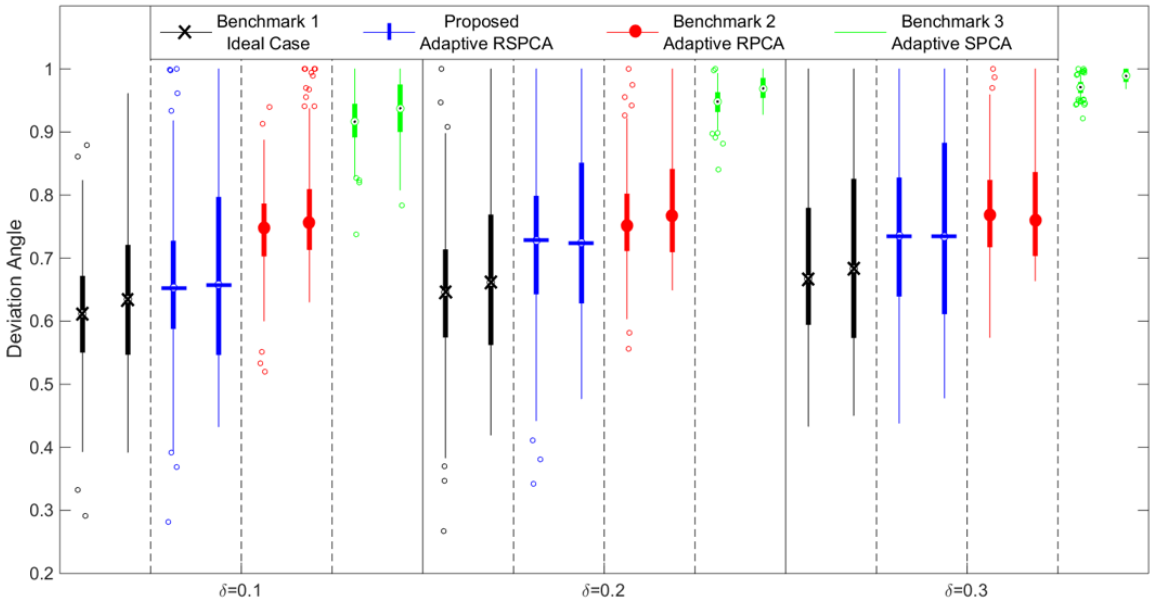


Figure 4.2 Box-plots of deviation angles for $n_{\text{transitional}} = 50$

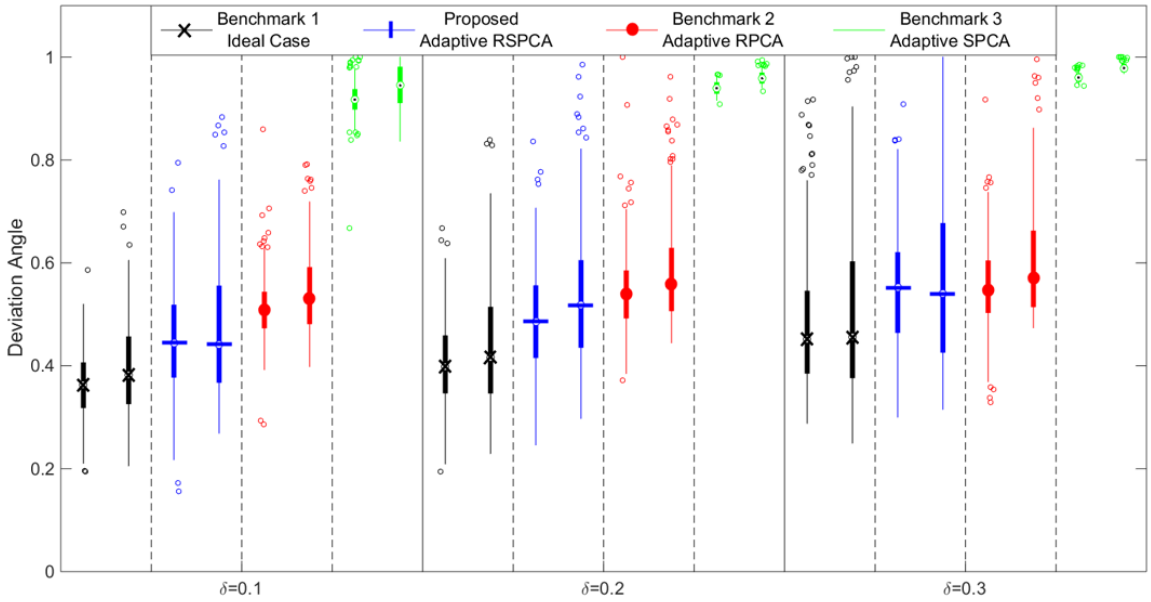


Figure 4.3 Box-plots of deviation angles for $n_{\text{transitional}} = 100$

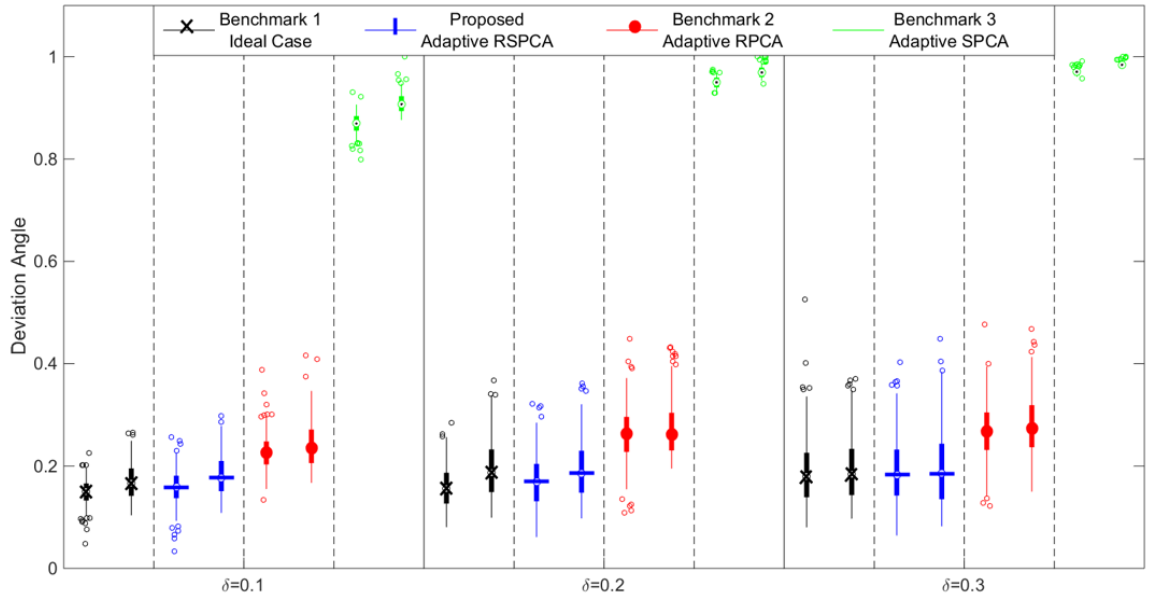


Figure 4.4 Box-plots of deviation angles for $n_{\text{transitional}} = 500$

4.5 Case Study

This section evaluates the online adaptive monitoring capability of our proposed method on a real case study of in-line Raman spectroscopy for (CNT) buckypaper manufacturing process. The data consists of 200 spectrums, where the first half is from the first setting and the out of control phase starts after spectrum 180. An initial model is fitted using the data from the first setting using RSPCA. The online monitoring starts by updating the initial model using the data acquired at a faster sampling frequency and higher intensity, which represents the second setting for the process. We aim to demonstrate how the adaptive RSPCA method can effectively adjust the initial model's parameters after the transition. Figure 4.1 illustrates the in control Raman spectrums obtained from the experiments, while Figure 4.5 demonstrates the means of the first and second operation settings and the out of control (OC) mean.

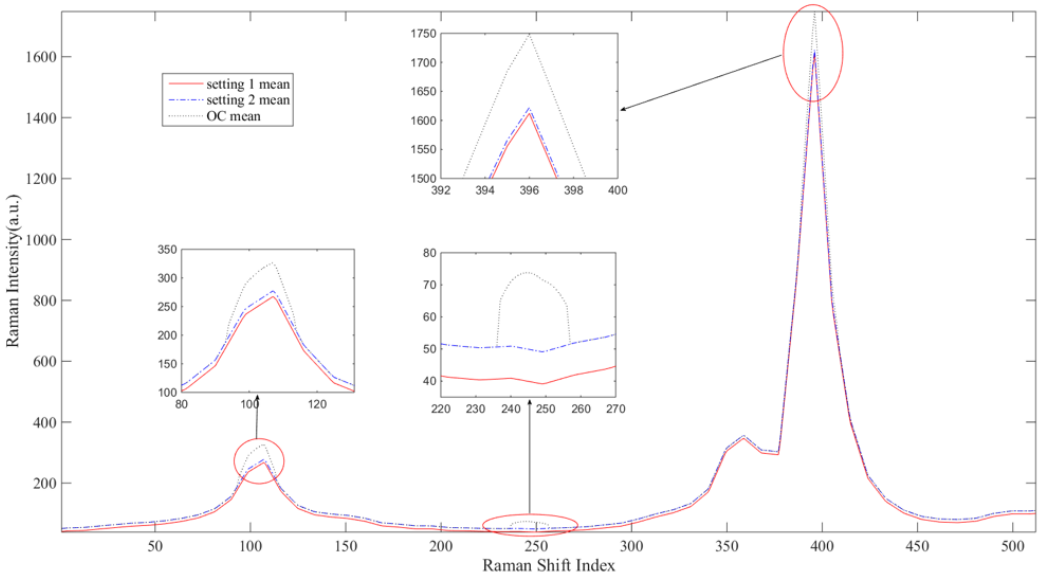


Figure 4.5 Out-of-control mean shift position and magnitude

In this study, we compare the monitoring performance of our method against sparse and regular variations of adaptive PCA (Liu et al. 2015b, Li et al. 2000). Additionally, we add the monitoring statistics obtained by using the initial model during the second setting. While an initial static model is generally expected to have poor monitoring performance for the new operating conditions, it can be used as a tool to detect the onset of the transition. It also serves to illustrate how an adequate adaptive procedure can significantly reduce the occurrences of false alarms due to the normal process transition.

The monitoring statistics are obtained from the first three leading principal components. The upper control limit (UCL) of The static and adaptive variations of robust sparse probabilistic PCA is set to $\chi^2_{0.99,3}$. While Li et al. (2000) suggests a time varying threshold, the threshold is still approximated by χ^2 but with different degrees of freedom depending on the number of chosen principal components. However, this will be constant in our case since we preset the number of components to three. For the adaptive sparse method we use kernel density estimation to determine the UCL as proposed by Liu et al. (2015b).

Figure 4.6 illustrates the monitoring statistics obtained using the different models for a representative run. The reported statistics start from spectrum 110, since the first 10 observations are used for the first batch update. As expected, the static RSPCA model estimated from the first process settings results in several false alarms during the online monitoring of the process after the transition. The other adaptive PCA and the sparse variant procedures also suffer from high false alarm rates, this is due to the lack of robustness in the models. In this representative run, all the benchmark methods detect the fault at time 193, which is a delay of 13 upon fault onset. Our procedure was capable of

signaling immediately at time 181 due to its accurate representation of the process parameters. This experiment was repeated 100 times by bootstrapping the Raman spectra. Table 4.2 summarizes the monitoring performance in terms of signaled false alarms and fault detection delays.

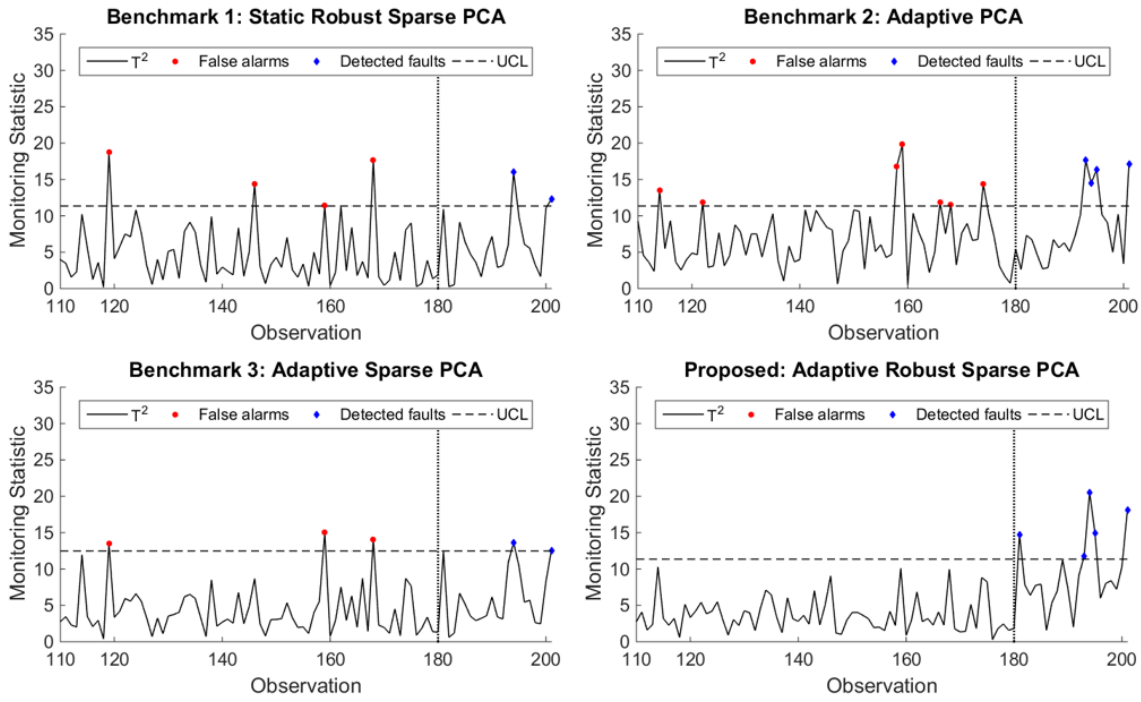


Figure 4.6 Monitoring performance of the proposed RSPCA compared to other benchmarks

Table 4.2 False alarm rates and detection delay comparison

Method	Benchmark 1 Static Robust Sparse PCA	Benchmark 2 Adaptive PCA	Benchmark 3 Adaptive Sparse PCA	Proposed Adaptive Robust Sparse PCA
False alarm rate	4.8%	7.75%	3.5%	*0.8%*
Mean detection delay (standard deviation)	7.2 (4.6)	4.2 (4.9)	9.5 (5.8)	*1.9 (1.2)*

4.6 Conclusion

Adaptive representation of manufacturing processes is crucial in dynamic manufacturing processes. This chapter proposes a novel method that adaptively finds robust estimates the spatial structure of sparse high dimensional data streams. This is achieved by using stochastic variational inference for solving the probabilistic robust sparse principal component analysis (RSPCA) model of the previous chapter. In addition, a robust adaptive learning rate is proposed to hedge against overfitting the model to novel observations that may be outliers. Moreover, the probabilistic approach in modelling retains the advantage of having an intuitive transition to process monitoring and diagnosis.

The performance of the proposed methodology was evaluated and compared to other adaptive PCA based techniques. The results from the simulation study demonstrates the effectiveness of our proposed procedure in addressing the issue of nonstationary operating conditions with sparse irregularities and outliers. The case study of inline Raman spectroscopy for (CNT) buckypaper manufacturing further illustrated the capability of the method to adaptively adjust the extracted features using novel observations from the process while simultaneously monitoring the incoming data streams.

Our proposed approach mainly relies on having complete access to all data streams for model estimation. A common issue in high dimensional settings is the inability to acquire or process full observations. Since the proposed method is based on a probabilistic model, it can be adjusted for implementation in the event of missing data. This may be of interest in many settings but it was not discussed in this chapter as it deserves its own analysis in future research.

CHAPTER 5. CONCLUSION

The development in the technology for observing key processes parameters using copious and diverse sensors has grown drastically in the past few years. This resulted in an abundance of information creating new ventures for advancement in process monitoring techniques. However, these endeavors carry along with them steep challenges that hinder current state of the art techniques to be ineffective. These data rich environments bear complex structures that may be sparse, noisy and full of outliers. In addition, monitoring limitations due to sensor unavailability or deployment cost can add another layer of complexity. These issues should be carefully explored and understood before attempting to implement any process monitoring technique. Naïve implementation of existing monitoring techniques that do not directly address these issues can result in misleading conclusions that can adversely affect the process.

In Chapter 2, we proposed a dimension reduction technique with the objective of having inherent monitoring and diagnostic capabilities embedded in the modelling procedure. Probabilistic models offer an intuitive platform for making inferences on model parameters. Specifically, we developed a probabilistic variation of the prevalently used principal component analysis (PCA) technique. Our proposed "Robust Sparse Principal Component Analysis" (RS-PCA) incorporates elements in the model that induce sparsity and promote robustness. Collectively, these two alterations allow for more accurate and consistent estimations of the extracted subspace. This was validated through simulated experiments and also tested on real data from in-line Raman spectroscopy.

In Chapter 3, we extend the work of Chapter 2 to time varying processes. We propose to adaptively update the probabilistic model via stochastic variational inference. Furthermore, we propose a robust adaptive learning rate that hedges against overfitting the model to novel observations that may be outliers. The performance of the proposed procedure was evaluated and compared to other adaptive PCA based techniques. The results from the simulation and case studies demonstrate the effectiveness of our proposed procedure in addressing the issue of nonstationary operating conditions with sparse irregularities and outliers.

In Chapter 4, we developed an adaptive sampling strategy that aims to circumvent limitations that commonly arise in data rich environments. We proposed a sampling strategy that is adaptive in the sense that it selects sensor to be observed online based on past observations as well as exploiting the embedded special structure. The developed Correlation based Dynamic Selection (CDS) technique uses an imputation of the observed stream to recreate a full spectrum for decision making rather than being limited by the partial stream. The method was validated by simulation and the performance was evaluated on the detection of solar flares from images.

The work presented in this thesis paves the way towards solving many other challenges that arose with the advent of modern sensing technology. The work of Chapters 2 and 3 on probabilistic PCA can be adjusted to address the common issue of missing data, due to the available probabilistic model. The ideas of the Chapter 4 may be extended to cases where the normality assumption may be too limiting. This can be achieved by exploring other nonparametric approaches to the compensation method, which was inspired by the

celebrated upper confidence bound (UAB) algorithm from the multi-armed bandit problem.

In summary, this dissertation aims to shed the light on opportunities available for improvement in process monitoring in data rich environments. The detrimental characteristics of such environments were individually highlighted and addressed in the chapters of this thesis and the performance of suggested methodologies was tested on realistic applications. That being said, high dimensional data can have several other nuances that can give rise to additional challenges which in return will necessitate a need for techniques capable of dealing with them.

APPENDIX A

In this Appendix, A.1 and A.2 provide the proofs for properties 1 and 2 of the proposed CDS algorithm, which were discussed in subsection 2.3.1.5. The following Lemma 1, which essentially follows from the weak law of large numbers, will be used in the proofs in A.1 and A.2.

Lemma 1: For an independent and identically sequence of a bivariate normal random variables $x_t \; and \; y_t$, such that $E[x] = \mu_x > \mu_y = E[y]$:

$$\lim_{T\to\infty} P\left(\sum_{t=t_0}^{T} x_t > \sum_{t=t_0}^{T} y_t\right) \to 1$$

Proof of lemma: Define the random variable $z_t = \sum_{t=t_0}^{T} x_t - \sum_{t=t_0}^{T} y_t$, then z_t is a Gaussian random walk. And we have:

$$\lim_{T\to\infty} P\left(\sum_{t=t_0}^{T} x_t > \sum_{t=t_0}^{T} y_t\right) = \lim_{T\to\infty} P(z_T > 0)$$

By assumption of $\mu_x > \mu_y$, we conclude that z_t is a random walk with a positive drift $E(x_t - y_t) = \mu_x - \mu_y > 0$, then it follows that:

$$\lim_{T\to\infty} (z_T) = \infty$$

and hence,

$$\lim_{T\to\infty} P(z_T > 0) = 1 \; \blacksquare$$

A.1 Proof of Property 1

At the high-level, the proof of property 1 involves two subcases. When there are no changes, the local statistics at sensors with infinitely many time observations will go back

to 0, whereas the local statistics at those local sensors without any observations and not correlated to observed sensors will be linearly increasing. Hence, we will sample from those non-observed sensors eventually. The second case is for when there is an insignificant change, where the linear increase of the unobserved sensors will still outrun the increase of the observed counterparts.

Since our sensor sampling procedure (algorithm 2.1) starts with picking elements of ω_t according to the $\max_i C_{i,t}$, it suffices to show that for any unobserved variable $x_{k'} \notin \omega_{t_0}$ there exists a time t such that $C_{k',t} > \max_k C_{k,t}$. If we take any unobserved variable $x_{k'} \notin \omega_{t_0}$ that is also not in the neighborhood of ω_{t_0} $\left(\text{i.e. } corr[x_{k'}, x_k] = 0 \text{ for all } x_k \in \omega_{t_0}\right)$, the increments of the positive and negative CUSUM will depend on $U_{k',t} = L_{k',t} = \Phi(1-\alpha/2)$. Then without loss of generality we can only consider the positive CUSUM $\left(C_{k'}^{+}\right)$. Hence, property 1 can be proven by comparing the increments of the CUSUM statistics from elements in C_k to those of $C_{k'}$, and showing that there exists a time t such that $C_{k',t} > \max_k C_{k,t}$. It suffices to show that there exists T such that when $\delta > 0$:

$$\sum_{t=t_0}^{T}\left(\delta U_{k',t} - \frac{\delta^2}{2}\right) > \sum_{t=t_0}^{T}\left(\delta x_{k,t} - \frac{\delta^2}{2}\right),$$

or equivalently,

$$z_T = \sum_{t=t_0}^{T}\left(U_{k',t} - x_{k,t}\right) > 0.$$

The assumption on $|E[x_k]|$ from property 1 can be broken down into two cases. First, we consider the case when $|E[x_k]| < \Phi(1-\alpha/2)$. Since $x_{k',t}$ is not in a neighborhood of $x_{k,t} \in \omega_{t_0}$, $E[U_{k'}] = \Phi(1-\alpha/2)$. Hence, z_T is a random walk with a positive drift and by Lemma 1:

$$P(z_T > 0) \to 1$$

The second case is when $E[x_k] = \Phi(1-\alpha/2) = E[U_{k'}]$. In this case, z_t becomes a Gaussian random walk with no drift. Let $H = inf\{z_t : t \geq 1\}$, then $H \overset{as}{\to} -\infty$ as $t \to \infty$ (Gut 1988). Hence, for any two variables $x_{k',t}$ and $x_{k,t}$ there exists a time t such that $C_{k',t} > C_{k,t}$. ■

A.2 Proof of Property 2

It suffices to show that increments of significantly out-of-control samples will be greater than the compensation given to the unobserved variables outside its neighborhood. Specifically, if we define $z'_t = \sum_{t=t_0}^{T}\left(x_{k,t} - U_{k',t}\right)$, and $|E[x_k]| > \Phi(1-\alpha/2)$ by the assumption in property 2, then z'_t is a random walk positive drift ($E[x_k] - E[U_{k'}]$). As $t \to \infty$ then $z'_t \to \infty$, this implies that there exists time t_0 such that $\forall t \geq t_0\ z'_t \geq 0$ and $C_{k',t} < C_{k,t}$. ■

It should be noted that the speed of the localization here depends on the drift $\left(E[x_k] \ - \Phi(1-\alpha/2)\right)$: the higher the post mean shift ($E[x_k]$) is, the faster it will diverge to ∞, which translates to quicker localization. Moreover, this shows that the sampling method will not favor a variable outside of the neighborhood. However, that does not mean that it will not explore the neighborhood even after it detects a faulty area. This essentially means that our method will not necessarily stick to the initial faulty area, but may still explore the surroundings to find an even bigger fault.

APPENDIX B

This Appendix provides the derivation of the adaptive learning rate in equation (4.13) of subsection 4.3.2. We begin by setting the objective to find a learning that minimizes the expected error $\Xi\left(\Theta^{*}_{(2),\mathrm{T}}, \Theta_{(2),\mathrm{T}}\right)$ as defined by equation 4.12. Substituting $\Theta_{(2),\mathrm{T}}$ with the expression given by equation (4.8) into the expected error yields the following:

$$E\left[\Xi\left(\Theta^{*}_{(2),\mathrm{T}}, \Theta_{(2),\mathrm{T}}\right)\right] = E\left[\left(\Theta^{*}_{(2),\mathrm{T}} - \Theta_{(2),\mathrm{T}-1} + \zeta_t\left(\Theta_{(2),\mathrm{T}-1} - \widehat{\Theta}_{(2),t}\right)\right)^{T}\right.$$
$$\left.\left(\Theta^{*}_{(2),\mathrm{T}} - \Theta_{(2),\mathrm{T}-1} + \zeta_t\left(\Theta_{(2),\mathrm{T}-1} - \widehat{\Theta}_{(2),t}\right)\right)\right].$$

Note that from subsection 4.3.2, the intermediate estimate of the global parameters $\widehat{\Theta}_{(2),t}$ is defined as a noisy estimate of the natural gradient such that $\mathrm{E}\left[\widehat{\Theta}_{(2),\mathrm{T}}\right] = \Theta^{*}_{(2),\mathrm{T}}$ and $\mathrm{Cov}\left[\widehat{\Theta}_{(2),\mathrm{T}}\right] = \mathrm{tr}(\Sigma_{\Theta})$. Hence, we can use these moments to compute the previous expectation to obtain the following expression:

$$E\left[\Xi\left(\Theta^{*}_{(2),\mathrm{T}}, \Theta_{(2),\mathrm{T}}\right)\right] = (1-\zeta_t)^2\left(\Theta_{(2),\mathrm{T}-1} - \Theta^{*}_{(2),\mathrm{T}}\right)^{T}\left(\Theta_{(2),\mathrm{T}-1} - \Theta^{*}_{(2),\mathrm{T}}\right) + \zeta_t^2\mathrm{tr}(\Sigma_{\Theta}).$$

Minimizing the above expression with respect to the learning rate ζ_t yields the result given by equation (4.13) of subsection 4.3.2.

REFERENCES

Archambeau, Cédric, and Francis R Bach. 2009. Sparse probabilistic projections. Paper read at Advances in neural information processing systems.

Archambeau, Cédric, Nicolas Delannay, and Michel Verleysen. 2006. Robust probabilistic projections. Paper read at Proceedings of the 23rd International conference on machine learning.

Archambeau, Cédric, Nicolas Delannay, and Michel Verleysen. 2008. "Mixtures of robust probabilistic principal component analyzers." *Neurocomputing* no. 71 (7-9):1274-1282.

Augusto, Carlos Roberto A, Anderson C Fauth, Carlos E Navia, Hisatake Shigeouka, and Kin H Tsui. 2011. "Connection among spacecrafts and ground level observations of small solar transient events." *Experimental Astronomy* no. 31 (2-3):177.

Banerjee, Onureena, Laurent El Ghaoui, and Alexandre d'Aspremont. 2008. "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data." *Journal of Machine learning research* no. 9 (Mar):485-516.

Ben-Gal, Irad, and Eugene Kagan. 2013. *Probabilistic search for tracking targets: Theory and modern applications*: John Wiley & Sons.

Cadima, Jorge, and Ian T Jolliffe. 1995. "Loading and correlations in the interpretation of principle compenents." *Journal of Applied Statistics* no. 22 (2):203-214.

Candès, Emmanuel J, Xiaodong Li, Yi Ma, and John Wright. 2011. "Robust principal component analysis?" *Journal of the ACM (JACM)* no. 58 (3):11.

Chatterjee, Snigdhansu, and Peihua Qiu. 2009. "Distribution-free cumulative sum control charts using bootstrap-based control limits." *The Annals of Applied Statistics*:349-369.

Chen, Tao, Elaine Martin, and Gary Montague. 2009. "Robust probabilistic PCA with missing data and contribution analysis for outlier detection." *Computational Statistics & Data Analysis* no. 53 (10):3706-3716.

Cheng, Qunfeng, Ben Wang, Chuck Zhang, and Zhiyong Liang. 2010. "Functionalized Carbon-Nanotube Sheet/Bismaleimide Nanocomposites: Mechanical and Electrical Performance Beyond Carbon-Fiber Composites." *Small* no. 6 (6):763-767.

Croux, Christophe, Peter Filzmoser, and Heinrich Fritz. 2013. "Robust sparse principal component analysis." *Technometrics* no. 55 (2):202-214.

Croux, Christophe, and Anne Ruiz-Gazen. 2005. "High breakdown estimators for principal components: the projection-pursuit approach revisited." *Journal of Multivariate Analysis* no. 95 (1):206-226.

d'Aspremont, Alexandre, Francis Bach, and Laurent El Ghaoui. 2008. "Optimal solutions for sparse principal component analysis." *Journal of Machine Learning Research* no. 9 (Jul):1269-1294.

De La Torre, Fernando, and Michael J Black. 2003. "A framework for robust subspace learning." *International Journal of Computer Vision* no. 54 (1):117-142.

Ding, Xinghao, Lihan He, and Lawrence Carin. 2011. "Bayesian robust principal component analysis." *IEEE Transactions on Image Processing* no. 20 (12):3419-3430.

Ding, Yu, Elsayed A Elsayed, Soundar Kumara, J-C Lu, Feng Niu, and Jianjun Shi. 2006. "Distributed sensing for quality and productivity improvements." *IEEE Transactions on automation science and engineering* no. 3 (4):344-359.

Duchi, John, Stephen Gould, and Daphne Koller. 2012. "Projected subgradient methods for learning sparse gaussians." *arXiv preprint arXiv:1206.3249*.

Efron, Bradley, and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap*: CRC press.

Feng, Jiashi, Huan Xu, and Shuicheng Yan. 2013. Online robust pca via stochastic optimization. Paper read at Advances in Neural Information Processing Systems.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics* no. 9 (3):432-441.

Frost, JR, and Lawrence D Stone. 2001. Review of search theory: advances and applications to search and rescue decision support. Soza and Company LTD Fairfax VA, Report No. CG-D-15-01.

Gao, Junbin. 2008. "Robust L1 principal component analysis and its Bayesian variational inference." *Neural computation* no. 20 (2):555-572.

Ge, Zhiqiang, and Zhihuan Song. 2011. "Robust monitoring and fault reconstruction based on variational inference component analysis." *Journal of Process Control* no. 21 (4):462-474.

George, Abraham P, and Warren B Powell. 2006. "Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming." *Machine learning* no. 65 (1):167-198.

Guan, Yue, and Jennifer G Dy. 2009. Sparse Probabilistic Principal Component Analysis. Paper read at AISTATS.

Gut, Allan. 1988. *Stopped Random Walks - Limit Theorems and Applications*. Applied Probability. A Series of the Applied Probability Trust, 5. Springer-Verlag, New York.

Han, Ningning, Yumeng Song, and Zhanjie Song. 2017. "Bayesian robust principal component analysis with structured sparse component." *Computational Statistics & Data Analysis* no. 109:144-158.

Hoffman, Matthew D, David M Blei, Chong Wang, and John Paisley. 2013. "Stochastic variational inference." *The Journal of Machine Learning Research* no. 14 (1):1303-1347.

Hsieh, Cho-Jui, Inderjit S Dhillon, Pradeep K Ravikumar, and Mátyás A Sustik. 2011. Sparse inverse covariance matrix estimation using quadratic approximation. Paper read at *Advances in neural information processing systems*.

Huber, Peter J. 1981. "Robust statistics." *Wiley Series in Probability and Mathematical Statistics, New York: Wiley,| c1981*.

Hubert, Mia, Tom Reynkens, Eric Schmitt, and Tim Verdonck. 2016. "Sparse PCA for high-dimensional data with outliers." *Technometrics* no. 58 (4):424-434.

Hubert, Mia, Peter J Rousseeuw, and Karlien Vanden Branden. 2005. "ROBPCA: a new approach to robust principal component analysis." *Technometrics* no. 47 (1):64-79.

Ishii, Takako T., Tomoko Kawate, Yoshikazu Nakatani, Satoshi Morita, Kiyoshi Ichimoto, and Satoshi Masuda. 2013. "High-Speed Imaging System for Solar-Flare Research at Hida Observatory." *Publications of the Astronomical Society of Japan* no. 65 (2). doi: 10.1093/pasj/65.2.39.

Jeng, Jyh-Cheng. 2010. "Adaptive process monitoring using efficient recursive PCA and moving window PCA algorithms." *Journal of the Taiwan Institute of Chemical Engineers* no. 41 (4):475-481.

Jin, Ran, Chia-Jung Chang, and Jianjun Shi. 2012. "Sequential measurement strategy for wafer geometric profile estimation." *Iie transactions* no. 44 (1):1-12.

Jolliffe, Ian. 2011. "Principal component analysis." In *International encyclopedia of statistical science*, 1094-1096. Springer.

Jolliffe, Ian T, Nickolay T Trendafilov, and Mudassir Uddin. 2003. "A modified principal component technique based on the LASSO." *Journal of computational and Graphical Statistics* no. 12 (3):531-547.

Kim, Dongsoon, and In-Beum Lee. 2003. "Process monitoring based on probabilistic PCA." *Chemometrics and intelligent laboratory systems* no. 67 (2):109-123.

Lai, Tze Leung. 1987. "Adaptive treatment allocation and the multi-armed bandit problem." *The Annals of Statistics* no. 15 (3):1091-1114.

Lai, Tze Leung, and Herbert Robbins. 1985. "Asymptotically efficient adaptive allocation rules." *Advances in applied mathematics* no. 6 (1):4-22.

Li, Guoying, and Zhonglian Chen. 1985. "Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo." *Journal of the American Statistical Association* no. 80 (391):759-766.

Li, Jing, and Jionghua Jin. 2010. "Optimal sensor allocation by integrating causal models and set-covering algorithms." *IIE Transactions* no. 42 (8):564-576.

Li, Weihua, H Henry Yue, Sergio Valle-Cervantes, and S Joe Qin. 2000. "Recursive PCA for adaptive process monitoring." *Journal of process control* no. 10 (5):471-486.

Li, Yongmin, L-Q Xu, Jason Morphett, and Richard Jacobs. 2003. An integrated algorithm of incremental and robust pca. Paper read at Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on.

Lim, Hock Beng, Mao Ching Foo, and Yulian Zeng. 2006. *An adaptive distributed resource allocation scheme for sensor networks*. Paper read at International Conference on Mobile Ad-Hoc and Sensor Networks.

Liu, Kaibo, Yajun Mei, and Jianjun Shi. 2015a. "An adaptive sampling strategy for online high-dimensional process monitoring." *Technometrics* no. 57 (3):305-319.

Liu, Kaibo, and Jianjun Shi. 2013. "Objective-oriented optimal sensor allocation strategy for process monitoring and diagnosis by multivariate analysis in a Bayesian network." *IIE Transactions* no. 45 (6):630-643.

Liu, Kangling, Zhengshun Fei, Boxuan Yue, Jun Liang, and Hai Lin. 2015b. "Adaptive sparse principal component analysis for enhanced process monitoring and fault isolation." *Chemometrics and Intelligent Laboratory Systems* no. 146:426-436.

Lois, Brian, and Namrata Vaswani. 2015. Online matrix completion and online robust pca. Paper read at Information Theory (ISIT), 2015 IEEE International Symposium on.

Ma, Zongming. 2013. "Sparse principal component analysis and iterative thresholding." *The Annals of Statistics* no. 41 (2):772-801.

Mandroli, Sampatraj S, Abhishek K Shrivastava, and Yu Ding. 2006. "A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes." *IIE Transactions* no. 38 (4):309-328.

Mason, Robert L, Nola D Tracy, and John C Young. 1995. "Decomposition of T2 for multivariate control chart interpretation." *Journal of quality technology* no. 27 (2):99-1108.

Mei, Yajun. 2010. "Efficient scalable schemes for monitoring a large number of data streams." *Biometrika* no. 97 (2):419-433.

Montgomery, Douglas C. 2009. *Introduction to Statistical Quality Control*: John Wiley & Sons (New York).

Nabhan, Mohammad, Yajun Mei, and Jianjun Shi. 2019. "HIGH DIMENSIONAL PROCESS MONITORING USING ROBUST SPARSE PROBABILISTIC PRINCIPAL COMPONENT ANALYSIS."

Parker, Eugene N. 1963. "The Solar-Flare Phenomenon and the Theory of Reconnection and Annihiliation of Magnetic Fields." *The Astrophysical Journal Supplement Series* no. 8:177.

Pereira, R Lopes, J Trindade, F Gonçalves, L Suresh, D Barbosa, and T Vazão. 2014. "A wireless sensor network for monitoring volcano-seismic signals." *Natural Hazards and Earth System Sciences* no. 14 (12):3123.

Pignatiello, Joseph J, and George C Runger. 1990. "Comparisons of multivariate CUSUM charts." *Journal of quality technology* no. 22 (3):173-186.

Portnoy, Ivan, Kevin Melendez, Horacio Pinzon, and Marco Sanjuan. 2016. "An improved weighted recursive PCA algorithm for adaptive fault detection." *Control Engineering Practice* no. 50:69-83.

Qiu, Chenlu, Namrata Vaswani, Brian Lois, and Leslie Hogben. 2014. "Recursive robust pca or recursive sparse recovery in large but structured noise." *IEEE Transactions on Information Theory* no. 60 (8):5007-5039.

Ranganath, Rajesh, Chong Wang, Blei David, and Eric Xing. 2013. An adaptive learning rate for stochastic variational inference. Paper read at International Conference on Machine Learning.

Rato, Tiago, Marco Reis, Eric Schmitt, Mia Hubert, and Bart De Ketelaere. 2016. "A systematic comparison of PCA-based Statistical Process Monitoring methods for high-dimensional, time-dependent Processes." *AIChE Journal* no. 62 (5):1478-1493.

Robbins, Herbert, and Sutton Monro. 1985. "A stochastic approximation method." In *Herbert Robbins Selected Papers*, 102-109. Springer.

Schaul, Tom, Sixin Zhang, and Yann LeCun. 2013. No more pesky learning rates. Paper read at International Conference on Machine Learning.

Scheinberg, Katya, Shiqian Ma, and Donald Goldfarb. 2010. Sparse inverse covariance selection via alternating linearization methods. Paper read at *Advances in neural information processing systems*.

Scheinberg, Katya, and Irina Rish. 2009. "SINCO-a greedy coordinate ascent method for sparse inverse covariance selection problem." *IBM Research Report* no. RC24837.

Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller. 1997. Kernel principal component analysis. Paper read at International Conference on Artificial Neural Networks.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*:267-288.

Tipping, Michael E, and Christopher M Bishop. 1999a. "Mixtures of probabilistic principal component analyzers." *Neural computation* no. 11 (2):443-482.

Tipping, Michael E, and Christopher M Bishop. 1999b. "Probabilistic principal component analysis." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* no. 61 (3):611-622.

Vidal, René, Yi Ma, and S Shankar Sastry. 2016. "Robust Principal Component Analysis." In *Generalized Principal Component Analysis*, 63-122. Springer.

Wainwright, Martin J, and Michael I Jordan. 2008. "Graphical models, exponential families, and variational inference." *Foundations and Trends® in Machine Learning* no. 1 (1–2):1-305.

Wang, A, X Xian, F Tsung, and K Liu. 2017. "A Spatial Adaptive Sampling Procedure for Online Monitoring of Big Data Streams." *Journal of Quality Technology*.

Wang, Chuan, Hsiao-Chun Wu, and Jose C Principe. 1996. Cost function for robust estimation of PCA. Paper read at Aerospace/Defense Sensing and Controls.

Wang, Xun, Uwe Kruger, and George W Irwin. 2005. "Process monitoring approach using fast moving window PCA." *Industrial & Engineering Chemistry Research* no. 44 (15):5691-5702.

Xian, Xiaochen, Rick Archibald, Benjamin Mayer, Kaibo Liu, and Jian Li. 2018a. "An effective online data monitoring and saving strategy for large-scale climate simulations." *Quality Technology & Quantitative Management*:1-17.

Xian, Xiaochen, Andi Wang, and Kaibo Liu. 2018b. "A nonparametric adaptive sampling strategy for online monitoring of big data streams." *Technometrics* no. 60 (1):14-25.

Xie, Yao, Jiaji Huang, and Rebecca Willett. 2013. "Change-point detection for high-dimensional time series with missing data." *IEEE Journal of Selected Topics in Signal Processing* no. 7 (1):12-27.

Yan, Hao, Kamran Paynabar, and Jianjun Shi. 2018. "Real-time monitoring of high-dimensional functional data streams via spatio-temporal smooth sparse decomposition." *Technometrics* no. 60 (2):181-197.

Yue, Xiaowei, Kan Wang, Hao Yan, Jin Gyu Park, Zhiyong Liang, Chuck Zhang, Ben Wang, and Jianjun Shi. 2017a. "Generalized wavelet shrinkage of inline Raman spectroscopy for quality monitoring of continuous manufacturing of carbon nanotube buckypaper." *IEEE Transactions on Automation Science and Engineering* no. 14 (1):196-207.

Yue, Xiaowei, Hao Yan, Jin Gyu Park, Zhiyong Liang, and Jianjun Shi. 2017b. "A Wavelet-Based Penalized Mixed-Effects Decomposition for Multichannel Profile Detection of In-Line Raman Spectroscopy." *IEEE Transactions on Automation Science and Engineering*.

Yue, Xiaowei, Hao Yan, Jin Gyu Park, Zhiyong Liang, and Jianjun Shi. 2018. "A Wavelet-Based Penalized Mixed-Effects Decomposition for Multichannel Profile Detection of In-Line Raman Spectroscopy." *IEEE Transactions on Automation Science and Engineering, 15(3), pp.1258-1271.*

Zeng, Jing, Kangling Liu, Weiping Huang, and Jun Liang. 2017. "Sparse probabilistic principal component analysis model for plant-wide process monitoring." *Korean Journal of Chemical Engineering* no. 34 (8):2135-2146.

Zhang, Zhengdao, Bican Peng, Linbo Xie, and Li Peng. 2015. "Process monitoring based on recursive probabilistic PCA for multi-mode process." *IFAC-Papers OnLine* no. 48 (8):1294-1299.

Zhang, Zhenyue, Hongyuan Zha, and Horst Simon. 2002. "Low-rank approximations with sparse factors I: Basic algorithms and error analysis." *SIAM Journal on Matrix Analysis and Applications* no. 23 (3):706-727.

Zhang, Zhenyue, Hongyuan Zha, and Horst Simon. 2004. "Low-rank approximations with sparse factors II: Penalized methods with discrete Newton-like iterations." *SIAM journal on matrix analysis and applications* no. 25 (4):901-920.

Zhu, Jinlin, Zhiqiang Ge, and Zhihuan Song. 2014. "Robust modeling of mixture probabilistic principal component analysis and process monitoring application." *AIChE journal* no. 60 (6):2143-2157.

Zoghi, M, and MH Kahaei. 2010. "Adaptive sensor selection in wireless sensor networks for target tracking." *IET Signal Processing* no. 4 (5):530-536.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. "Sparse principal component analysis." *Journal of computational and graphical statistics* no. 15 (2):265-286.